The Dissertation Committee for Leslie Keng certifies that this is the approved version of the following dissertation:

A Comparison of the Performance of Testlet-Based Computer Adaptive Tests and Multistage Tests

COMMITTEE:

Barbara G. Dodd, Supervisor

S. Natasha Beretvas

Randall Parker

Keenan Pituch

Tiffany Whittaker

A Comparison of the Performance of Testlet-Based Computer Adaptive Tests and Multistage Tests

by

Leslie Keng, B. Math, B. Ed, M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2008

Dedicated to my family

Acknowledgements

As the popular saying goes, “It takes a village to raise a child”. Being a father of two young (and adorable) girls, I certainly agree with this. However, my experience in completing this dissertation also leads me to believe that, “It takes a village to complete a doctorate”. And I hope this acknowledgement does justice to the “village” of people that have helped me get to this point.

I want to express extreme gratitude to my amazing dissertation advisor, mentor, and friend, Dr. Barbara Dodd. Barbara took me under her wings from day one and has helped me every step along the way. From making sure that I would always be able to support my young family, to developing me as a disciple and researcher in the field of psychometrics, and providing me with the opportunity to work for a testing company, Barbara has truly gone above and beyond what a typical advisor does for her students. And needless to say, I would not have finished my dissertation at the rate I did without her encouragement and confidence in me. Thank you, Barbara – any accomplishment that I will ever have in this field would not be possible without your inspiration and guidance.

I am very grateful to Dr. Tasha Beretvas, who I like to (jokingly) say, got me into this mess in the first place. Little did I know that when, as a master student in the math stats program, I took an out-of-department course in the educational psychology department that I would end up spending my next four years and the foreseeable future in this field, and I am very grateful that I will be. This field is truly a wonderful marriage of my interests and background in math, statistics, education and computer science. And I would not be here had Tasha not pointed it out to me – frequently and persistently, I might add. Thank you, Tasha. I hope that this is not the end, but the continuation of even more collaborations between us in the future.

I would also like to thank the rest of my dissertation committee members, Dr. Randy Parker, Dr. Keenan Pituch and Dr. Tiffany Whittaker, for bearing with me throughout the entire dissertation process. Their suggestions and over-generous compliments were very helpful in shaping this dissertation.

I have been blessed with a wonderfully supportive group of people that I will have the privilege of continuing to work with after the completion of my doctorate. The Texas psychometric services team at Pearson has been so an encouraging to me during this entire process and I want to thank each and every one of them from the bottom of my heart. I especially would like to thank my mentors and supervisors at Pearson, Dr. Laurie Davis and Dr. Kimberly O'Malley, for allowing me to work for Pearson over the past few years as a psychometric intern. I have learned so much for them, not only about practicing good psychometrics and the testing industry, but also about how to be great managers who lead by example.

Next, I would like to thank my church family at Austin Chinese Church (ACC). Being a foreigner with no immediate family close by, this dear group of brothers and sisters at ACC has been my extended family in Austin for the past 10 years. I became a Christian because of their love and support when I first moved here, and that by itself is the most valuable gift anyone could receive. On top of that, I cannot even count the number of brothers and sisters that have encouraged and generously supported me over the years, both financially and through prayer. May the Lord bless each and every one of you for what you have done for the least of His brothers (Matthew 25:40).

Throughout my life, I have been blessed with loving family members, and I would like to dedicate this dissertation to my family:

To my grandfather, whom I wish could physically be here to share in this accomplishment, but I trust is smiling as he looks down from heaven.

To my father, Gene Keng, who has inspired me to not only follow in his footsteps in my career, but has also taught me by example how to be a man of integrity.

To my brother, Ken, who is my best man, my confidant, and the person I took things out on before I got married.

To my girls, Faith and Carissa, who are likely too young right now to understand what this is all about, but Daddy wants you to know that you both inspired me to finish this. And I hope and pray that some day this will inspire you to follow your dreams and see it to completion.

My greatest blessings, however, come from two amazing women who took on, as their careers, the most difficult but most important job in the world: stay-at-home moms.

To my mother, Liza Keng, who spent over 25 years of her life raising my brother and me, who has been there for everything, including all of our ceremonies, big and small, and who is one of the main reasons our entire family will be spending eternity in heaven. Thank you! This accomplishment is as much mine as it is yours.

And to my beloved wife, Flora, who is my biggest helper, my best friend, and my soul mate. We have shared this journey every step of the way, through the agonies, frustrations, tears (if I could cry) and joy. Ani, your dedication to everything you do has inspired me beyond what you probably realize. I hope that one day, when you find something you want to pursue, that I can be as supportive and inspiring to you as you were to me. If I could add your name next to mine on this dissertation and my degree, I would. Thank you and love you, always!

Last, but not least, I want to give thanks to the Lord Jesus Christ for Your saving grace and faithfulness, despite my frequent lack of faith. You have shown me time and again that You are faithful and true to Your Word when you said, “I will be with you; I will never leave you nor forsake you” (Joshua 1:5). I pray that I will continue to follow

the path you have set forth for me, growing in faith, love, and hope, and deepening in my personal relationship with you – the most precious thing any person can have.

April 29, 2008

A Comparison of the Performance of Testlet-Based Computer Adaptive Tests and Multistage Tests

Publication No. _____

Leslie Keng, Ph. D.

The University of Texas at Austin, 2008

Supervisor: Barbara G. Dodd

Computer adaptive testing (CAT) has grown both in research and implementation. Test construction and security issues, however, have led many to reconsider the merits of CAT. Multistage testing (MST) is an alternative adaptive test design that purportedly addresses CAT's shortcomings. Yet considerably less research has been conducted on MST. Also, most research in adaptive testing has been based on item response theory (IRT). Many tests now make use of testlets – bundles of items administered together, often based on a common stimulus. The use of testlets violates local independence, a fundamental assumptions of IRT. Testlet response theory (TRT) is a relatively new measurement model designed to measure testlet-based tests. Few studies though have examined its use in testlet-based CAT and MST designs.

This dissertation investigated the performance of testlet-based CATs and MSTs measured using the TRT model. The test designs compared included a CAT that is adaptive at the testlet level only (testlet-level CAT), a CAT that is adaptive at both the

testlet and item levels (item-level CAT) and a MST design (MST). Test conditions manipulated included test length, item pool size, and examinee ability distribution. Examinee data were generated using TRT-calibrated item parameters based on data from a large-scale reading assessment. The three test designs were evaluated based on measurement effectiveness and exposure control properties.

The study found that all three adaptive test designs yielded similar and good measurement accuracy. Overall, the item-level CAT produced better measurement precision, followed by the MST design. However, the MST and CAT designs yielded better measurement precision at different areas of the ability scale. All three test designs yielded acceptable exposure control properties at the testlet level. At the item level, the testlet-level CAT produced the best overall result. The item-level CAT had less than ideal pool utilization, but was able to meet its pre-specified maximum exposure control rate and maintain low item exposure rates. The MST had excellent pool utilization, but a higher percentage of items with high exposure rates. Skewing the underlying ability distribution also had a particularly notable negative effect on the exposure control properties of the MST.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE: INTRODUCTION

The movement towards computer-based testing as a viable alternative to traditional paper-and-pencil testing has gradually become a reality in large-scale educational assessments. This movement has been brought about not only by advancements in computing technology, but also by key developments in modern test theory. Computer-based testing allows for flexibility in the time and place examinees are administered their tests, and it facilitates more accurate and rapid reporting of test results to a large number of examinees (Bergstorm & Lunz, 1999). It also makes possible the administration of the assessments in ways other than the traditional linear fixed format in which all examinees get identical sets of items. One such alternative computer-based test design is a computer adaptive test (CAT).

In a CAT, each examinee receives a tailored test with a set of items that most closely matches their estimated proficiency or ability level. The administration of a CAT is analogous to the test-giving approach of an intelligent human test administrator who takes into account how the examinee has performed so far on the items given, and chooses items that he or she believes are most accessible to the examinee. Thus, one of the main advantages of CATs is that it can shorten the length of the test while still achieving equivalent or better measurement precision of the examinee's ability (Weiss, 1982; Wainer, 2000). Consequently, computer adaptive testing has become a popular mode of administration in recent decades. Many large-scale educational testing programs and licensure and credential agencies offered CAT versions of their assessments.

The implementation of a CAT design was made possible by the development of the family of measurement models known as item response theory (IRT; Rasch, 1960; Lord & Novick, 1968). IRT describes, in mathematical terms, the relationship between

an examinee's ability and the probability of a given response to a test item based on characteristics of the item. It overcomes many of the shortcomings of classical true score theory (Gulliksen, 1950), putting item characteristics and examinee ability on the same scale, thereby allowing the examinee's proficiency to be related to his or her performance at the item level instead of only at the test score level. This is an important attribute of IRT because it enables the administration of different sets of items to different examinees while still being able to estimate their abilities on the same scale (Embretson & Reise, 2000). This attribute enables the creation of testing algorithms that construct individualized tests for examinees, the essential feature of a CAT. IRT models differ in the number of item parameters they assume and the type of items they measure. Dichotomous IRT models are used to measure tests whose item responses can be classified into binary categories, such as multiple-choice items. The commonly-used dichotomous IRT models include the one-, two- and three-parameter logistic models (1PL, 2PL and 3PL respectively; Rasch, 1960; Birnbaum, 1968). Polytomous IRT models, on the other hand, can be applied to tests whose items have more than two response categories. A number of polytomous IRT have been proposed and implemented (e.g. Samejima, 1969; Andrich, 1978; Masters, 1982). Both dichotomous and polytomous IRT models have been applied to CATs in a variety of testing contexts.

IRT, however, is not without its limitations. One item format that has become increasingly popular because of its efficiency in both item development and administration is the testlet. A testlet is defined as a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow (Wainer & Kiely, 1987). Common examples in practice include a set of reading items associated with a common passage or a set of mathematics items referencing a single table or graphic. The use of testlets poses a

challenge to IRT because of the fundamental assumption of local independence in IRT. Local independence means that, conditional on an examinee's ability, the probability of correctly responding to an item is statistically independent of the probability of responding correctly to any other item (Hambleton & Swaminathan, 1985). Item responses within a testlet, however, are not entirely independent; they are related through the common stimulus. Using IRT to measure a test consisting of testlets can thus lead to inaccurate estimation of examinee and item parameters and overestimation of the precision of these parameters (Tuerlinckx & De Boeck 2001; Sireci, Thissen & Wainer 1991).

A number of approaches have been suggested to address the issue of local dependency in testlet-based tests. One common approach is to define the testlet as the unit of measurement and then apply one of the polytomous IRT models (Wainer & Lewis, 1990). Under this approach, a testlet becomes the unit of measurement and is viewed as a single polytomous item with possible scores ranging from zero to the total number of items in the testlet. While this approach has been shown to work well in an array of situations (Wainer, 1995), there are two scenarios where it falls short. One is when more information needs to be extracted from the item response patterns within the testlet, and the other is when ad hoc testlet construction is desired in the context of CAT (Wainer, Bradlow & Wang, 2007). With the polytomous IRT approach, response patterns that lead to the same total score for a testlet cannot be distinguished and the items associated with a testlet must be fixed for all examinees. A more recently proposed alternative approach to modeling testlets that can handle these scenarios is testlet response theory (TRT).

In TRT, the item remains the unit of measurement and an additional parameter is included with the dichotomous IRT models to account for the shared variance among

items within a testlet (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow & Du, 2000). Two- and three-parameter TRT models have been proposed and are available for use in a variety of testing situation, such as in testlet-based CATs. The TRT models allow a testlet-based CAT to be administered in two different ways. One way is for the CAT to only be adaptive between testlets. That is, the CAT chooses each testlet it administers based on the examinee's estimated ability. However, once a testlet is selected, the associated items that each examinee receives are linear and fixed. Such a CAT is termed a *testlet-level CAT* in this dissertation. To date, this is the way that testlet-based CATs are typically administered (e.g. Boyd, 2003). The other way TRT allows a testlet-based CAT to be given is by adapting not only between testlets, but also within each testlet. Thus, not only is the testlet to be administered selected based on the examinee's estimated ability, but also the associated items that an examinee receives within the testlet. Such a CAT is called an *item-level CAT* for this study. While the ability to model a testlet-based item-level CAT is one of the purported advantages of TRT over the polytomous IRT approach, no research to date has examined the gains in psychometric properties for a item-level CAT measured with TRT.

Also, the use of CAT as a test design has come under much scrutiny in recent years due in large part to test security issues (Chang, 2004). Because, at each step in a CAT, the testing algorithm chooses the most informative item based on the examinee's estimated ability level, examinees with similar abilities can end up with significant overlap in the sets of items they are given over the entire test. Additionally, since little is known about each examinee at the start of a CAT, a default ability estimate, such as the mean of the assumed ability distribution, is often assumed for each examinee, resulting in the same items being given as the first few items of the test to the majority of examinees. These two properties of the CAT testing algorithm result in unbalanced utilization of the

item pool, with a small set of items being administered to most of the examinees (Hulin, Drasgow & Parsons, 1983). These items tend to have high exposure rates, which can lead to security breaches that threaten the integrity of the test if examinees share information about the test with one another (e.g. Davey & Nering, 2002). As such, an item or testlet exposure control procedure typically needs to be implemented with the CAT testing algorithm. Several exposure control procedures have been proposed (e.g. McBride & Martin, 1983; Sympson & Hetter, 1985; Chang & Ying, 1996) and compared in recent years (e.g. Pastor, Dodd & Chang, 2002; Davis & Dodd, 2003). A recent study (Boyd, 2003) found that the progressive-restrictive exposure control procedure (Revuelta & Ponsoda, 1998) was ideal for controlling the exposure rates of testlet-based CATs modeled with TRT.

CAT has also been criticized for some of its shortcomings related to test administration. One such criticism is the lack of administrative control over the content quality of a CAT. Every CAT is built on-the-fly during test administration for each examinee. Thus, it cannot be reviewed a priori by test developers, particularly content specialists, to ensure all content requirements are met and that no context effects exist, such as one item cuing the answer to another. Even with the most sophisticated content balancing algorithm, it is generally not possible to code every content specification in the test blueprint. Thus, the use of human review for test quality assurance is often an essential step in the test development process (Luecht & Nungester, 1998), but it is difficult to implement in a CAT. Another criticism of CAT is the lack of review opportunities for examinees. Most CATs prohibit examinees from skipping items or reviewing and editing previous item responses during the test (Lunz & Bergstrom, 1994; Vispoel, 1998). This inflexibility exists to prevent examinees from using test-taking strategies that would circumvent the testing algorithm and threaten the measurement

efficiency of the test. However, examinees find this to be a big disadvantage and it is one of the most common complaints about CAT from test takers (Patsula, 1999). Because of these administrative shortcomings of a CAT, an alternative computer-based test design known as multistage testing has been proposed and implemented by several testing programs in recent years.

A multistage test (MST) can be viewed as a middle ground between the traditional linear fixed format test and CAT (Jodoin, Zenisky & Hambleton, 2006). It is adaptive, thus it can generally achieve greater measurement precision than a linear fixed format test (Luecht, Nungester & Hadadi, 1996). Its points of adaptation, however, are not between items, but between pre-assembled bundles of items known as modules. The modules are arranged into a set number of test stages with pre-determined routing rules specifying how an examinee can move from one stage to the next based on their current performance on the test (Luecht & Nungester, 1998). This gives test developers, such as content specialists, more administrative control over a MST compared to a CAT, allowing them to review the test content for quality assurance prior to its administration. Multiple forms, called panels, of a MST are usually built and randomly selected for administration to each examinee, thereby controlling for the exposure and utilization rates of modules and the items and testlets within each module (Jodoin, 2003).

While the concept of a multistage test is not a new one (e.g. Cronbach & Glaser, 1965; Lord 1971, 1974), theoretical development and practical implementation of MST have only begun in earnest recently due to the issues related to CAT and the movement towards computer-based testing in many testing programs. As such, the implications of various MST test design considerations, such as test length, item pool size and the assumed underlying ability used to build the test, are still relatively unknown and are the topics of ongoing MST research. And while it has been suggested in MST literature

(Zenisky, 2004; Hendrickson, 2007), no study to date has examined the use of TRT as a measurement model for testlet-based MSTs.

Several studies have compared the performance of MST with CAT (e.g. Kim & Plake, 1993; Luecht et al., 1996; Patsula, 1999; Jodoin, 2003; Davis & Dodd, 2003). The general finding has been that CATs can achieve better measurement precision across the full range of ability levels, but the psychometric advantages need to be weighed against the greater administrative control afford by MSTs. Only a handful of studies, however, have compared the exposure control properties of MSTs and CATs. This is an important issue related to test security that requires further investigation if multistage testing is to become a viable alternative computer-based test design to computer adaptive testing in high-stakes large-scale assessment programs.

The purpose of this dissertation, therefore, was to compare the measurement effectiveness and exposure control properties of two testlet-based CAT designs and one MST design across several manipulated test conditions. The two testlet-based CAT designs included the testlet-level CAT and the item-level CAT, each implementing the progressive-restrictive exposure control procedure; while an eight-panel three-stage MST with one module in the first stage and three modules each in the second and third stages (known as the 1-3-3 stage structure) were constructed. All three designs were measured with the three-parameter logistic (3PL) TRT model. The manipulated test conditions included the total test length, the size of the item pool, and the underlying population distribution from which examinee abilities were sampled. Testlet and item parameters in the item pool were based on real data from recent administrations of a statewide reading examination. And realistic CAT and MST simulations were conducted to compare the merits and limitations of each design in terms of measurement accuracy and precision as well as several exposure control indices.

CHAPTER TWO: LITERATURE REVIEW

The literature review in this chapter provides background information related to the proposed dissertation study. It contains five main sections. The first section provides an introduction to item response theory, including its assumptions, characteristics and the different types of item response theory models. One limitation of item response theory is its difficulty in modeling local dependency in tests that contain item bundles known as testlets. The second section, therefore, describes an extension of item response theory called testlet response theory, designed specifically to handle such scenarios. It gives a description of the Bayesian statistics framework on which testlet response theory is built, followed by details about the different types of testlet response theory models. The third section discusses the popular test administration framework known as computer adaptive testing. It provides details about the various components of a computer adaptive test, along with research pertaining to computer adaptive testing and its various issues. The fourth section introduces an alternative testing framework called multistage testing. It introduces the components of a multistage test, followed by details of multistage testing design considerations along and related research in literature. The section also gives an overview of the studies that have compared computer adaptive testing with multistage testing. The final section is the statement of problem, summarizing the issues and research questions that are of interest in this dissertation.

ITEM RESPONSE THEORY

Item Response Theory (IRT; Rasch, 1960; Lord & Novick, 1968) is a family of mathematical models used to describe the relationship between the probabilities of a given response to an item conditional on an examinee's ability level. The fundamental

IRT equation expressing this relationship is known as an item characteristic function or item characteristic curve (ICC).

IRT was derived as an extension to classical true score theory (Gulliksen, 1950). One important limitation of classical true score theory is its inability to separate the dependency between examinee characteristics and test characteristics – each must be interpreted in the context of the other. IRT overcomes this shortcoming through its property of *parameter invariance*. That is, the estimated ability of an examinee does not depend on the set of items the examinee is administered; and the estimated characteristics of an item do not depend on the particular set of examinees to which it is given. This and several other more theoretically justifiable measurement principles have increased the popularity of IRT in both research and operational test settings over the past few decades (Embretson & Reise, 2000). The application of IRT principles is especially prevalent in the areas of large-scale assessments and computerized adaptive tests.

Assumptions of IRT

Three basic assumptions underlie IRT models. The first assumption is that, given an examinee's ability level, it is possible to find a mathematical function to describe the probability of a given examinee response to an item based on the item's characteristics (Hambleton & Swaminathan, 1985). Additional assumptions can be made about the item characteristics, such as item difficulty, discrimination power, and guessing probability, that affect the shape of the ICC. Different assumptions about item characteristics result in different measurement models under IRT.

The second assumption is that a single ability or trait is measured by the set of items that make up the test. This is commonly referred to as *unidimensionality* (Hambleton & Swaminathan, 1985). For example, if IRT is used to model a reading comprehension test, then it is assumed that any statistical dependency among the item

responses is accounted for by the examinee's reading ability. Extensions have been made to IRT so that a set of test items can measure multiple abilities or traits (Reckase, 1997). However, multidimensional IRT is not examined in this dissertation and will therefore not be discussed further.

The third assumption of IRT is *local independence* (Hambleton & Swaminathan, 1985). It means that conditional on the examinee's ability, the probability of responding to an item is statistically independent of the probability of responding to any other item. Mathematically, local independence means that the probability of an examinee's item response pattern is equal to the product of the individual probabilities from each item's ICC at the examinee's ability level. The assumption of local independence has been shown to be equivalent to that of unidimensionality (Lord, 1980; Lord & Novick, 1968). Together, they underscore the key IRT property that the only factor affecting an examinee's responses to a set of test items is the examinee's ability of interest. This has several important implications to test items constructed using IRT principles. For example, it means that the content of one item should not give any clues to the answer of another item on the test. Also, the set of test items should also not be related through some common stimulus such as a passage or prompt.

Types of IRT Models

Numerous IRT models have been proposed since IRT was first described. These models can be classified in general as *dichotomous* and *polytomous* IRT models.

Dichotomous IRT models are typically associated with tests that have multiple-choice items, but can also be applied to any test whose item responses can be classified into binary categories, such as correct and incorrect, true and false, or agree and disagree (Embretson & Reise, 2000). These item types can be scored accurately and efficiently, and hence appear in a variety of large-scale educational and psychological assessments.

Consequently, the dichotomous IRT models are the most commonly employed IRT model in practical applications. The three most prevalent dichotomous IRT models include the one parameter logistic (1PL; Rasch, 1960; Wright & Stone, 1970), the two parameter logistic (2PL; Birnbaum, 1968) and the three parameter logistic (3PL; Birnbaum, 1968) IRT models.

Polytomous IRT models are often applied to tests consisting of items with multiple ordered response categories. For example, a reading test may have constructive response or essay items that are not scored simply as correct or incorrect, but are scored on a scale of say, 1 to 5. Multiple-category items are especially prevalent in measurement instruments within the attitude and personality assessment domains (Embretson & Reise, 2000) and they are generally more informative than dichotomously-scored items. Most polytomous IRT models were derived as extensions to one of the dichotomous IRT models and simplify to them when an item has only two score categories. Examples of polytomous IRT models include the graded response model (Samejima, 1969) and the modified graded response model (Muraki, 1990), the nominal response model (Bock, 1972), the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), the successive interval model (Rost, 1988) and the rating scale model (Andrich, 1978). Because polytomous IRT models are not examined or analyzed in this dissertation, specific details of these polytomous IRT models will not be given.

Dichotomous IRT Parameters

The three dichotomous IRT models are characterized by the number of item parameters included in the model. The simplest model is the 1PL IRT model (or the Rasch model) where items are distinguished by how difficult they are. Thus, each item has only one item parameter, b, known as the *item difficulty* parameter. The 2PL IRT model allows items to vary not only in difficulty, but also in discrimination power.

Thus, in addition the item difficulty parameter (*b*), it also includes a parameter, *a*, to indicate *item discrimination*. The 3PL IRT model further extends the simpler models, recognizing that even an uninformed examinee can get an item correct by chance; that is, each item may be answered correctly through guessing. So in addition to item discrimination (*a*) and difficulty (*b*), the 3PL IRT model incorporates a *pseudo-guessing* parameter (*c*).

To illustrate the attributes of the three IRT item parameters, consider how they are defined in relation to the following item characteristic curve (ICC) for an item described under the 3PL IRT model (in Figure 2.1). For dichotomous items, the ICC is a graphical representation showing the probability of answering a given item correctly conditional on the examinee's ability level, denoted as θ (Hambleton, Swaminathan & Rogers, 1991). For the item illustrated in Figure 2.1, the difficulty parameter, *b*, is equal to -0.5; the discrimination parameter, *a*, equals 1.2; and the pseudo-guessing parameter, *c*, is 0.2.

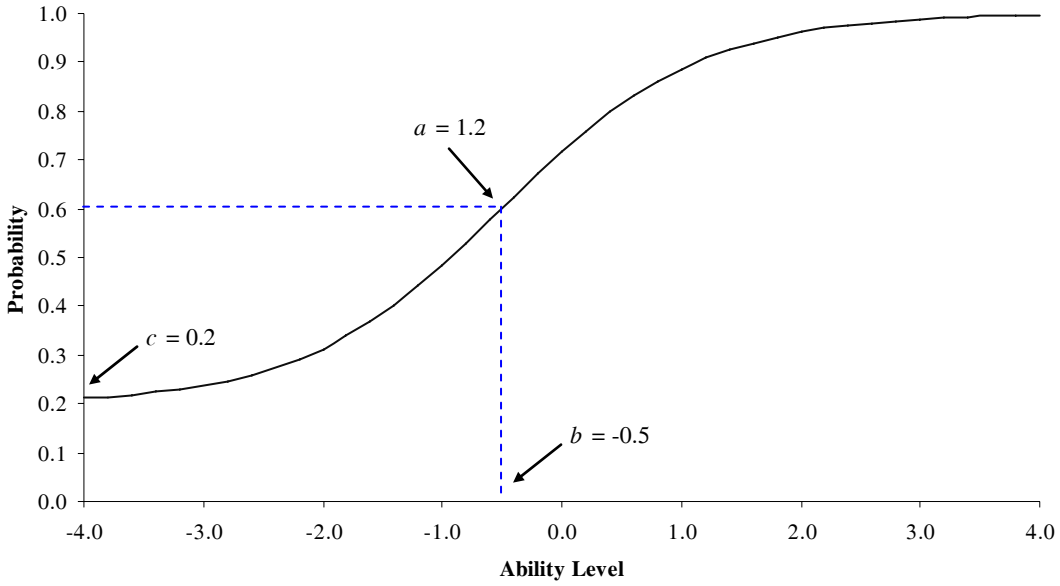

Figure 2.1: Item Characteristic Curve for a 3PL IRT item

The difficulty parameter (*b*) indicates the relative difficulty or easiness of the item. The value of *b* is equal to the ability scale (θ) value that corresponds to the point of inflection for the ICC; that is, the point on the ICC where the slope is maximal (see Figure 2.1). It is known as a *location* parameter because the value of *b* locates the position of the ICC in relation to the ability scale (Hambleton, Swaminathan & Rogers, 1991). Thus, items that are more difficult have larger *b* values, and their ICCs are therefore located further to the right on the ability scale. The range of the difficulty parameter is from -∞ to +∞, but the *b* values for most items are typically between -3 to +3.

The discrimination parameter (*a*) identifies how well an item can distinguish between examinees that are more proficient in the measured trait (i.e. those with high θ values) from those who are less proficient. An item with a high *a* value is more discriminating and is more useful for separating examinees into different ability levels. In terms of the ICC, the value of *a* is proportional to the slope of the ICC at the point of inflection. Thus, items with steeper slopes are more discriminating. The theoretical range of the discrimination parameter is (-∞, +∞). However, a negative discrimination value generally indicates a problem with the item, for instance miskeying, and such an item should be discarded. Thus, the *a* values for most items are positive and typically less than +2 (Hambleton, Swaminathan & Rogers, 1991).

The pseudo-guessing parameter (*c*) accounts for performance of examinees at the low end of the ability scale. For such examinees, even if they are not proficient enough to correctly answer the item, they still may get the item right by chance through guessing. Items can vary in how easy they are to guess. As such, *c* can range from 0 (no guessing possible) to +1, although they are typically closer to 0. As Figure 2.1 shows, the value of *c* corresponds to the lower asymptote of the ICC.

Item and Test Information

The value of an item's parameters affects the measurement precision of an examinee's ability level (θ). Under IRT, the precision of measurement for θ is not the same across the ability scale. Measurement precision is quantified with an item's information function, denoted I(θ) and expressed as,

$$I(\theta) = \frac{P'(\theta)^2}{P(\theta)(1-P(\theta))} \quad (1)$$

where $P(\theta)$ is the probability of an examinee correctly answering the item conditional on θ, and $P'(\theta)$ is the first derivative of $P(\theta)$ with respect to θ (Embretson & Reise, 2000). The higher the information function at a particular θ value, the more precisely the item can measure examinees with abilities at that level.

Formula (1) shows that an item information is related to the first derivative of $P(\theta)$, which corresponds to the slope of the item's ICC. Thus, the amount of information an item can provide is closely related to the item's discrimination (*a*) parameter. The more discriminating an item is, the more information it provides in measuring ability levels around the difficult parameter (*b*) value (Embretson & Reise, 2000).

Because of the assumption of local independence, information functions for items on a test built according to IRT principles are also independent and can be summed to attain the *test information function*, $TI(\theta)$,

$$TI(\theta) = \sum I(\theta) \quad (2)$$

Test information can be used to evaluate the measurement precision of a test. This is done by examining the *standard error* of θ, $SE(\theta)$, at the various ability levels. The standard error of θ is related to test information through the formula,

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad (3)$$

Note from formulas (2) and (3) that test information and standard error, and therefore measurement precision of θ, is not constant across the ability scale. Many tests are built such that the standard errors tend to be higher at the extremes and lower near the middle of the ability continuum (Embretson & Reise, 2000).

TESTLET RESPONSE THEORY

A *testlet* (Wainer & Kiely, 1987) or item bundle (Rosenbaum, 1988) refers to a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow. Examples include a set of reading comprehension items associated with a common passage or a set of mathematics items referencing a single word problem.

From a test development perspective, the use of testlets can help increase testing efficiency (Wainer, Bradlow & Wang, 2007). This is especially prevalent in situations where the goal is to assess an examinee's understanding of a stimulus; such as, a literary piece, a case study, a map, a musical passage, or a table of numbers. In these cases, examinees need a substantial amount of time to process the stimulus before answering any test items. It would therefore be an inefficient use of testing time and examinee effort if only a single item were associated with each stimulus.

Local Dependency in Testlets

The use of testlets, however, violates local independence, one of the basic IRT assumptions. This is because item responses within a testlet are not entirely independent – they are related through the common stimulus (Rosenbaum, 1988). Violation of local independence has been shown to lead to inaccurate estimation of item and person parameters (Tuerlinckx & De Boeck 2001; Chen & Thissen 1997; Ackerman 1987) and overestimation of test information functions and reliability (Sireci, Thissen & Wainer

1991; Thissen, Steinberg & Mooney, 1989). It also introduces additional dimensions (Wainer & Thissen, 1996), which violates unidimensionality, another basic IRT assumption.

One common approach have to handling local dependency in testlet data is to define the testlet as the unit of measurement and then apply one of the polytomous IRT models (Wainer & Lewis, 1990). For example, suppose we have a testlet with seven reading items associated with a common passage. Instead of calibrating each item individually using a dichotomous IRT model, this approach views the entire set of seven items as one single "item", scores it out of a maximum total of seven points, then calibrates it using one of the polytomous IRT model. This approach has been shown to work well (Wainer, 1995) in a board array of situations and is considered a practical solution to handling local dependency in testlets. However, there are two scenarios in which this approach falls short (Wainer et al, 2007).

One scenario is when we need to extract more information for the item response patterns within the testlet. Using the polytomous IRT approach, the testlet score is represented by the total number of correct items. While this representation is often sufficient, some information can be extracted for the exact patterns of correct responses. Continuing our example of the 7-item testlet, suppose two examinees each correctly answered 3 of the 7 reading items, one of the examinees, however, achieved this by answering the first 3 items correctly; while the other did so by answering the last 3 items correctly. Clearly, more information about the items and each examinee can be extracted from these response patterns. Under the polytomous IRT approach, however, the responses of these two examinees would be scored exactly the same.

Another scenario occurs in the context of computer adaptive tests (CAT) in which *ad hoc testlet construction* is desired. That is, in the spirit of maximizing item

information for each examinee taking a CAT, one may want to build the items within a testlet on the fly. For example, suppose that a total of 15 items in the item bank are associated with a particular reading passage. However, the intent was never to administer all 15 items to any given examinee. Instead, the CAT algorithm should adaptively select a subset of these items depending on each examinee's estimated ability and other constraints. Consequently, each examinee gets a different set of items within a testlet, and, if a variable-length CAT is permitted, each examinee may even get a different number of items for a testlet. Thus, the polytomous IRT approach cannot be applied in this scenario as the total number of correct items within a testlet no longer carries the same meaning from one examinee to another. A more complex model is required to handle these two scenarios involving testlets.

Testlet Response Theory

Testlet Response Theory (TRT; Wainer, Bradlow & Du, 2000) is another approach for modeling tests involving testlets. Unlike the polytomous IRT approach, TRT maintains the *item* as the unit of measurement and explicitly accounts for the local dependency between items within a testlet. As a result, it is able to handle the two scenarios mentioned above. Specifically, because TRT uses the item as the unit of measurement, it is possible to distinguish and extract information out of the item response pattern given by examinees within a testlet. In addition, because local dependency within testlets are estimated for each testlet and for each examinee, all examinees are not required to be administered the same set of items or even the same number of items within a testlet. Thus, ad hoc test construction is feasible under TRT.

TRT is embedded in a Bayesian framework. That is, it is implemented within a full probability model in which a joint probability distribution of all observable and unobservable quantities is provided. One benefit of the Bayesian approach is that it

allows us to capture our knowledge of the underlying test structure as well as the way that the data were collected. Such knowledge is modeled in what is known as a *prior distribution* (or simply, a *prior*), which is described for all the parameters of interest and incorporated into the probability model. Another advantage of the Bayesian approach is that, rather than just getting a single point estimate for any parameter of interest, we obtain the entire distribution for each parameter, known as the *posterior distribution* (or simply, a *posterior*). One can then sample from the posterior distribution to compute typical summary statistics such as means, standard deviations, and interval estimates, as well as to test any hypotheses. This is unlike in the traditional frequentist approach where only point estimates of parameters are obtained. Additional distributional assumptions, such as normality for measurement error, then need to be made in order to compute interval estimates and perform hypothesis tests.

The drawback of the Bayesian approach, however, is in the complexity of attaining the posterior distributions for the various parameters. A substantial amount of computing power is required to perform Bayesian statistical analyses. These obstacles have been overcome in recent years with the advent of computing technology in conjunction with the development of Markov Chain Monte Carlo (MCMC) estimation methods such as the Metropolis-Hasting algorithm (Metropolis, Rosenblith, Rosenblith, Teller & Teller, 1953; Hasting, 1970) and the Gibbs sampler (Geman & Geman, 1984). A plethora of computational methods and application (e.g. Tanner & Wong, 1987; Gelfand and Smith, 1990; Albert, 1992; Albert & Chib, 1993) have been proposed and implemented based on MCMC that have made the Bayesian approach feasible and extendable to a variety of contexts, such as educational and psychological measurement. TRT was hence developed by Bradlow, Wainer & Wang (1999) under the Bayesian framework.

Dichotomous Testlet Response Theory Models

Bradlow et al. (1999) have developed the Bayesian TRT models as extensions to the traditional IRT models. Thus, there are the TRT-equivalent of the dichotomous IRT models known as the two-parameter logistic (2PL) and three parameter logistic (3PL) TRT models. In addition, a TRT model that can be applied to tests with a *mixture* of binary and polytomous testlet data has been proposed (Wang, Bradlow & Wainer, 2002). Extensions have also been made to all the TRT models so that the parameters can be a function of *covariates* (Wainer et al., 2007). The inclusion of covariates can help answer some of the *why* questions related to the test, such as why certain items were more difficult or why a group of students excelled on the test. In this dissertation, only the dichotomous TRT models are examined. Thus, no further discussion is given on the mixture TRT models and TRT models with covariates. Those interested can consult Wainer et al. (2007) for more information.

The 2PL Testlet Response Theory Model

The two parameter logistic testlet response theory model (2PL-TRT) is the initial testlet model developed by Bradlow et al. (1999). It serves as a baseline to the more complex TRT models and is a modification of the 2PL IRT model (Birnbaum, 1968) to account for local dependency between items within the same testlet. For the 2PL-TRT model, the probability of person i with ability level, θ_i, correctly answering item j within testlet, $d(j)$, is denoted as $P_{ij}(y_j = 1 \mid \theta_i)$ and is expressed as,

$$P_{ij}(y_j = 1 \mid \theta_i) = \frac{\exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})}{1 + \exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})} \qquad (4)$$

where, as in the traditional 2PL IRT model, the parameters a_j and b_j represent the item discrimination and item difficulty respectively for item j. The additional parameter, $\gamma_{id(j)}$,

is called the *testlet effect parameter* and models the extra dependency for person *i* responding to item *j* nested within testlet *d(j)*.

Note that the testlet effect parameter, $\gamma_{id(j)}$, is both a person and testlet parameter. This means that for a given testlet, *d(j)*, the effect of the local dependency of the testlet items varies for each examinee. Thus, the *variance* of $\gamma_{id(j)}$ is typically estimated for each testlet and used as an indicator of the degree of local dependency within each testlet.

The 2PL-TRT is embedded within a Bayesian framework, which allows the sharing of information across examinees, items, and testlets (Wainer et al, 2007). Thus, prior distributions for each of the model parameters need to be specified. They include,

1. $\theta_i \sim N(0,1)$
2. $\log(a_j) \sim N(\mu_a, \sigma_a^2)$
3. $b_j \sim N(\mu_b, \sigma_b^2)$
4. $\gamma_{id(j)} \sim N(0, \sigma_\gamma^2)$

The means and variance components in these prior specifications (that is, μ_a, σ_a^2, μ_b, etc) are also given slightly informative normal and inverse-gamma priors (called *hyperpriors*) respectively to ensure that the posterior distributions can be properly estimated. Note that for the 2PL-TRT, a single parameter, σ_γ^2, is assumed for the variance for the testlet effect parameter, $\gamma_{id(j)}$, across all testlets. This implies that the degree of extra local dependence due to the testlet effect is the same for every testlet.

Bradlow et al. (1999) demonstrated the efficacies of the 2PL-TRT model and the Bayesian computational method through an extensive simulation study comparing three measurement models. The compared models included the 2PL IRT model analyzed using BILOG (Mislevy & Bock, 1983) with the traditional frequentist approach, the 2PL IRT model embedded within the Bayesian framework analyzed with an MCMC algorithm (Tanner & Wong, 1987), and the 2PL-TRT model analyzed also with the same

MCMC sampling algorithm. The three models were simulated on two different testlet size conditions (that is, number of items within each testlet) crossed with three different testlet effect conditions (that is, the amount of testlet effect variance, σ_γ^2), and a baseline condition where no testlet effect was assumed. The results showed that the MCMC approach applied to 2PL IRT performed equivalently to the BILOG approach in terms of mean absolute errors for θ, a and b, and 95% coverage probability of θ. And while the three models performed similarly on the baseline (no testlet effect) condition, the 2PL-TRT MCMC model outperformed the other two methods as the testlet size and the testlet effect increased. It yielded the lowest mean absolute errors for θ, a and b, and the most accurate 95% nominal coverage for θ in the presence of local dependency.

The 3PL Testlet Response Theory Model

The Bayesian three parameter logistic testlet response theory model (3PL-TRT) was proposed by Wainer et al.(2000) as an analog for the standard 3PL model in IRT (Birnbaum, 1968) as well as an extension to the 2PL-TRT. It is analogous to the 3PL IRT model in that in incorporates an additional item parameter, c_j, for guessing. It extends the 2PL-TRT not only by being able to handle guessing in binary data, but also so that different testlets may exhibit substantially different amounts of local dependency. Thus, a single parameter to describe the variance of the testlet effect (σ_γ^2) is no longer assumed in the 3PL-TRT. For the 3PL-TRT, the probability of person i with ability level, θ_i, correctly answering item j within testlet, $d(j)$, is expressed as,

$$P_{ij}(y_j = 1 \mid \theta_i) = c_j + (1 - c_j) \cdot \frac{\exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})}{1 + \exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})} \qquad (5)$$

where a_j and b_j again represent item discrimination and item difficulty respectively for item j; and c_j is the pseudo-guessing parameter for item j. The testlet effect parameter,

$\gamma_{id(j)}$, still models the extra dependency for person *i* responding to item *j* nested within testlet *d(j)*.

The prior distributions for θ_i, a_j, and b_j are the same as the 2PL-TRT, but the prior for $\gamma_{id(j)}$ is changed to $\gamma_{id(j)} \sim N(0, \sigma^2_{d(j)})$. In addition, the prior distribution for the pseudo-guessing parameter is specified as, $\log(\frac{c_j}{1-c_j}) \sim N(\mu_c, \sigma_c^2)$. Slightly informative normal and inverse-gamma hyperpriors are again given for the means and variance components (that is, μ_a, σ_a^2, μ_b, etc) respectively. Note that the variance for the testlet effect parameter is no longer restricted to be the same for every testlets, but is a testlet-specific variance component.

Wainer et al.(2000) conducted a simulation study to compare the performance of the 3PL-TRT to 3PL IRT models. The four models compared included the 3PL IRT model estimated using marginal maximum likelihood (MML), the 3PL IRT model estimated with MCMC, the 3PL-TRT model with a common testlet effect variance, σ^2_γ (abbreviated as $MCMC_\gamma$) and the 3PL-TRT model with testlet-specific testlet effect variances, $\sigma^2_{d(j)}$ (abbreviated as $MCMC_d$). The data were simulated with three testlet effect conditions – no testlet effect, equal testlet effects (across testlets), and unequal testlet effects. The results showed that all models performed similarly in recovering the true parameter values when no testlet effect was present; although the three models using MCMC estimation outperformed the one using MLL in recovering the *a* and *c* parameters. However, when testlet effects were present, the two 3PL-TRT models performed better than the two IRT models in parameter recovery. Additionally, $MCMC_d$ outperformed $MCMC_\gamma$ in the unequal testlet effects condition. These results demonstrate the efficacy of the 3PL-TRT in modeling local dependency, particularly in situations where the degree of dependency varies across testlets.

Wainer et al. (2000) also applied the four models in their simulation study to real data sets from the Scholastic Assessment Test (SAT) and the Graduate Record Examination (GRE). The verbal sections of each of these tests contained a number of testlets. For the SAT, the amount of testlet effect variance estimated by the TRT models was quite small. Thus, the four models fit similarly to the SAT data. For the GRE, however, the variance of the testlet effect was found to be substantially larger. Consequently, the IRT models produced very different parameter estimates than the TRT models. Most notably, the IRT models found significantly larger estimates for the item discrimination parameters (a_j). This finding is important because item discrimination is closely related to the item information function, which for the 3PL-TRT is given by (Wainer et al., 2000),

$$I(\theta_i) = a_j^2 \cdot \left(\frac{\exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})}{1 + \exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})} \right)^2 \cdot \frac{1 - c_j}{c_j + \exp(a_j \cdot (\theta_i - b_j - \gamma_{id(j)})} \qquad (6)$$

Thus, by ignoring the local dependency due to the testlet effects, the IRT models overestimate a_j. This in turn inflates item information and leads to standard errors that are too small (Wainer et al., 2007). Thus, by more appropriately modeling for local dependency, the 3PL-TRT gives a more accurate account of the amount of measurement precision in the GRE verbal test.

COMPUTER ADAPTIVE TESTING

Computerized adaptive testing (CAT) is recognized as an efficient alternative mode of test administration to traditional paper-and-pencil (P&P) tests. Unlike paper-and-pencil tests (P&P) where all examines receive an identical set of items; CAT tailors the test for each examinee. It administers test items that more closely match each examinee's estimated ability. Consequently, the major advantage of CAT is that it can

shorten test length while increasing measurement precision of the examinee's ability (Weiss, 1982; Wainer, 2000).

In addition, as with all computer based tests (CBTs), a CAT system leverages the benefits of computer technology and leads to several advantages for examinees and test administrators such as flexibility in scheduling, increased testing opportunities, automated data collection, and prompt score reporting (Bergstorm & Lunz, 1999). As a result, CAT has become a popular mode of administration in recent decades. Several large-scale educational assessments such as the Armed Services Vocational Aptitude Battery (ASVAB), the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), and the National Council of State Boards of Nursing are administered as CATs (Chang, 2004).

Any CAT system consists of four main components: the item pool, item selection procedure, ability estimation method, and stopping rule (Reckase, 1989). These four components are described in the following sections.

Item Pool

An item pool is a large collection of items that may be administered to examinees on a CAT. Traditional P&P tests are also built from an item pool. P&P tests, however, differ from CATs in that any item that is selected from the pool for a P&P test remains constant for the given form of that test. For a CAT, each examinee gets an individualized test consisting of varying sets of items drawn from the pool. Thus, the quality of the item pool has a significant effect on the performance of the adaptive algorithm in a CAT (Flaugher, 2000). The quality of an item pool is based not only on the size of the pool, but also on the breadth of content coverage, depth of items in each content area, and the psychometrics properties of the items (Parshall, Spray, Kalohn & Davey, 2002).

Several considerations for developing a CAT item pool are similar to those for a P&P item pool. Items in both types of pools need to have been written according to content specifications, reviewed for content quality and test sensitivity, and pre-tested so that their psychometric properties can be evaluated before they are placed into the pool (Flaugher, 2000). Methods used to evaluate the quality of an item can combine traditional item statistics such as proportion correct and biserial correlation with IRT- or TRT-based criteria such as item parameters and item information (Wainer, 1989). Another important item characteristic to examine is whether each item fits the underlying measurement model, such as IRT or TRT. One method for assessing whether an item fits a model is the *analysis of item-ability regressions* (Kingston & Dorans, 1985). This graphical method compares the plot of the item's empirical frequency distribution conditional on examinee ability (θ) to the item's ICC based on its estimated item parameters. If the two plots are similar, then the item has good model fit and is appropriate for inclusion into the pool.

Some item pool considerations are particularly important for a CAT. Because items are adaptively selected to match each examinee's estimated ability, the item pool must contain high-quality items for several different levels of proficiency. In contrast, typical P&P tests are built with items that best measure average examinees; that is, those with proficiency levels near the center of ability distribution (Flaugher, 2000). As such, a CAT item pool tends to be larger than a P&P item pool so that different combinations of test items can be generated for a wide range of examinee abilities (Davey & Nering, 2002). The purpose of the CAT also influences the overall distribution of item information in the pool across the ability scale, that is, the pool information function. For a norm-referenced test (NRT), where the purpose is to measure trait levels equally well across the ability scale, the ideal shape for the pool information function is a rectangular

distribution (Reckase, 1981; Urry, 1977). On the other hand, for a criterion-referenced test (CRT), where the purpose is to measure the examinees' trait levels with respect to one or more points or cut scores along the ability scale, the pool information function should ideally peak at the cut scores (Parshall et al., 2002).

Item Selection Procedures

In a traditional P&P test, items are selected by the test constructor *prior* to the administration of the test. The distinguishing mark of a CAT is that it administers items that are most appropriate for each examinee. As a result, items are chosen for administration *during* the test as more information is learned about each examinee. To accomplish this, a *testing algorithm* is needed that selects test items adaptively. One of the first adaptive testing algorithms was the *flexilevel test* described by Lord (1971). The flexilevel test algorithm was an adaptive test design that was not computerized nor based on any complex measurement models such as IRT or TRT. Over the years, however, more sophisticated testing algorithm have been developed that incorporate statistical procedures and modern measurement models while leveraging advancements in computing technology. Two commonly used item selection algorithms are: *maximum information* and *Bayesian* selection (Thissen & Mislevy, 2000; Kingsbury & Zara, 1989).

Maximum Information Selection

The goal of most item selection procedures is to accumulate as much test information, $TI(\theta)$, as possible in the most efficient manner (Parshall et al., 2002). As seen from formula (3), as the amount of test information increases, measurement precision of the ability estimate also increases. The maximum information (MI) selection procedure chooses, at each step of the CAT, an item from the pool that provides the maximum amount of item information, $I(\theta)$, given the provisional estimate of the

examinee's ability, θ (Lord, 1977; Brown & Weiss, 1977). By maximizing the incremental information provided with each item, the MI procedure is also maximizing the expected precision of θ and doing so with substantially less items than traditional P&P tests.

One issue with MI selection relates to a problem in test theory known as the *attenuation paradox* (Lord & Novick, 1968). This paradox typically occurs at the beginning of a CAT because the errors in the initial ability estimates are generally large. Thus, items that perform best at the provisional ability estimate may in fact perform worse at the true ability value (van der Linden & Pashley, 2000). This problem is exacerbated by MI selection because it generally chooses items with high discrimination power. Such items tend to have peaked information functions, providing maximal information over a narrow ability range around the inaccurate provisional ability estimate, while providing minimal information outside that range where the examinee's true ability may lie (Parshall et al., 2002).

Bayesian Selection

A Bayesian counterpart to the MI procedure is known as the maximum posterior precision selection procedure (Owen, 1969, 1975). This procedure, at each step, chooses the item that maximizes the precision of the posterior ability distribution. This procedure overcomes the issue of large errors in the provisional ability estimates, especially at the beginning at a CAT, by selecting items based on the entire posterior ability distribution instead of a single point estimate. Thus, while the selected item may not provide maximum information at the provisional ability estimate, it is the most informative on average across the high density region of the posterior distribution (Parshall et al., 2002). The disadvantages of this procedure, however, include that is it can be far more

computationally intensive than MI and that the ability estimate is sensitive to the order in which items are administered (Thissen & Mislevy, 2000).

Running either the MI or Bayesian selection procedures as described above (that is, unconstrained) typically leads to undesired patterns of item usage. The most notable patterns include overexposure of certain items in the pool, particularly items with high discrimination power (*a* values) and unbalanced administration of items from the various content areas (Thissen & Mislevy, 2000). Exposure control and content balancing are important issues that the testing algorithm needs to address to ensure test security and content validity of a CAT. Thus, these issues will be discussed in detail in a later section.

Level of Selection

An additional decision related to the testing algorithm when a CAT consists of testlets is the level of selection. The most common level of selection with testlet-based CATs is at the *testlet* level. This means when a testlet is chosen, based on whichever selection criteria, all items that are predefined to be associated with the testlet are administered (Wainer & Kiely, 1987). In other words, such an algorithm is adaptive at the testlet level, but is linear and fixed at the item level.

An alternative level of selection in testlet-based CATs is to adapt at the *item* level. Thus, within a chosen testlet, items are administered adaptively such that each examinee can take a different set of items. It is unclear whether adapting at the item level within testlets provides more precise ability estimates (Thissen & Mislevy, 2000). A few studies have shown that the gain in precision is only modest when adapting within testlets (Wainer, Lewis, Kaplan & Braswell, 1991; Wainer, Kaplan & Lewis, 1992). However, these studies were conducted prior to the invention of testlet response theory. They examined tests with only a single testlet and hence very short test lengths. They were also not conducted in the context of CAT, but were done on *hierarchical* testlets, which

are testlets with predefined tree structures containing all the possible paths the examinees could take based on their item responses. One of the purported advantages of TRT is that it permits *ad hoc testlet construction* in the context of a CAT (Wainer, Bradlow & Du, 2000; Wainer et al., 2007). However, no study to date has evaluated the gain in efficiency under such a scenario using this new measurement model for testlets.

Ability Estimation

In most CAT systems, the parameter values for items in the pool are assumed to have been pre-tested and calibrated before the items are administered operationally. Thus, the only parameter that requires estimation during CAT administration is the examinee's proficiency or ability level, θ.

The first step in ability estimation process involves determining an initial ability estimate. The initial ability estimate is needed for the testing algorithm to choose the first item or testlet on the test. One way to determine the *initial* ability estimate is to use prior information known about the examinee, such as the examinee's previous test scores in the same subject area. Or, it can simply be set to the mean of the assumed distribution, which would be zero, if θ is assumed to be from the standard normal distribution.

After each item or testlet is given, interim or *provisional* estimates of θ are typically needed by the CAT algorithm to choose the next item or testlet. The *final* ability estimation is then performed at the end of the test based on the examinee's entire set of responses. The provisional and final ability estimates do not have to be obtained using the same method (Chang, Ansley & Lin, 2000). The final ability estimate may also be transformed to a different ability metric (Parshall et al., 2002). Two common approaches to ability estimation in CAT are maximum likelihood estimation and Bayesian estimation.

Maximum Likelihood Estimation

A likelihood function, $L(\theta)$, describes the probability of observing the set of item responses, $\mathbf{Y} = (y_1, y_2, \ldots, y_k)$, where y_j is the examinee's response to item j given that the examinee's ability level is θ. That is, the value of $L(\theta)$ for a specific value of θ represents the relative likelihood that $\mathbf{Y}$ would be observed if θ were the true examinee ability. A general equation for the likelihood function is given by,

$$L(\theta_i) \equiv P(\mathbf{Y} \mid \theta_i) = \prod_j P_{ij}(\theta_i)^{y_{ij}} Q_{ij}(\theta_i)^{1-y_{ij}} \tag{7}$$

where $P_{ij}(\theta_i)$ is the conditional probability of examinee i answering item j correct (that is, $y_{ij} = 1$) given that the examinee's ability level is θ_i; while $Q_{ij}(\theta_i) = 1 - P_{ij}(\theta_i)$. The expressions for $P_{ij}(\theta_i)$ and $Q_{ij}(\theta_i)$ are determined by the choice of measurement model. For example, for the 3PL-TRT, Equation (5) would specify the expression for $P_{ij}(\theta_i)$ and $Q_{ij}(\theta_i)$ (Wainer et al., 2000). Note that in (7), the product of conditional probabilities across the set of item responses only holds under the fundamental IRT assumption of local independence (Hambleton, Swaminathan & Rogers, 1991).

The maximum likelihood estimate (MLE) of θ is simply the mode of the likelihood function. One common way of finding the maximum of $L(\theta)$ is by first computing the log-likelihood function, that is, $log[L(\theta)]$ and finding its derivative,

$$\frac{\partial log[L(\theta)]}{\partial \theta} = \sum_j (y_{ij} - P_{ij}(\theta_i)) \frac{P'_{ij}(\theta_i)}{P_{ij}(\theta_i) Q_{ij}(\theta_i)} \tag{8}$$

Then, by setting (8) equal to zero, the MLE is the θ_i value that satisfies the equation. This equation is often solved iteratively, using a method such as the Newton-Raphson procedure (Embretson & Reise, 2000). The Newton-Raphson procedure is a popular numerical analysis method for approximating the roots of real-valued functions. It is an iterative algorithm that finds the roots using the first derivative on the function.

The measurement precision of the MLE, or $SE(\theta)$, is approximated as the square root of the reciprocal of the test information function, $TI(\theta)$, at the given θ estimate. Equation (2) gives the formula for the test information function. Note that in the context of CAT, because the test is being constructed adaptively, test information is increasing with the administration of each item. Because of the additive property of item information, the contribution to the precision of estimation of θ can be calculated for every administered item as well as for any potential items in the item pool (Wainer & Mislevy, 2000).

The MLE method has the advantage of being relatively unbiased compare to the Bayesian methods. However, it is unstable for short tests and can even be unbounded. This occurs, for example, when an examinee either answers all item correctly or all items incorrectly. In such cases, the likelihood function does not have a mode and the MLE would be $\pm\infty$ (Parshall et al., 2002). This can be particularly problematic at the start of a CAT where it is likely for an examinee's responses to be in one category (all correct or all incorrect). Therefore, an alternative method, such as one of the Bayesian procedures (described later), is often used at the beginning of a CAT until a stable ability estimate can be obtained with MLE. An additional issue with MLE is when the likelihood function has multiple modes or a number of local extrema. In such cases, solving the zeros for Equation (8) does not necessarily yield the global maximum, but only a local maximum or even a local minimum of the likelihood function. This issue is often resolved by carefully choosing a starting value for the Newton-Raphson procedure that would allow the iterations to locate the global maximum more easily instead of getting "stuck" on a local maximum or minimum (Wainer & Mislevy, 2000).

Bayesian Estimation

The distinguishing characteristics of Bayesian ability estimation procedures is the use of a prior distribution, $p(\theta)$, in conjunction with the likelihood function, $L(\theta)$. Recall that the prior distribution represents what is known about the distribution of θ before the test is administered. Thus, Bayesian procedures allow the incorporation of prior knowledge about the testing population into the ability estimation, enabling more efficient estimation (Embretson & Reise, 2000).

The Bayesian approach follows from a fundamental rule in probability theory known as *Bayes Theorem*, which relates the conditional probability of two events through their marginal probabilities (Barnard & Bayes, 1958). Applying Bayes Theorem to the CAT context gives the following mathematical relationship,

$$P(\theta \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \theta) \cdot p(\theta) = L(\theta) \cdot p(\theta) \tag{9}$$

The term $P(\theta|\mathbf{Y})$ in Equation (9) is known as the *posterior distribution* of θ. Bayesian procedures estimate examinee abilities by computing measures of central tendency for the posterior distribution. The *expected a posteriori* (EAP; Bock & Mislevy, 1982) estimator is the mean of the posterior distribution; while the *maximum a posteriori* (MAP) estimator is the mode.

Computationally, MAP estimation works very similarly to MLE. The MAP estimate for a given set of item responses, **Y**, can be obtained by first computing the derivative of the log-posterior distribution with respect to θ. That is,

$$\frac{\partial log[P(\theta \mid \mathbf{Y})]}{\partial \theta} = \frac{\partial log[L(\theta)]}{\partial \theta} + \frac{\partial log[\, p(\theta)]}{\partial \theta} \tag{10}$$

The solution to Equation (10) can then be found using an iterative process such as the Newton-Raphson procedure. Measurement precision of the MAP ability estimate can be approximated by,

$$SE(\theta) \approx \left[\sqrt{TI(\theta) - \frac{\partial^2 p(\theta)}{\partial \theta^2}} \right]^{-1} \tag{11}$$

where $TI(\theta)$ is the test information function. Notice that in Equation (11), $TI(\theta)$ is augmented with a term based on the second derivative of the prior distribution. This term is usually negative. As such, the measurement precision for the MAP ability estimate typically exceeds that of the MLE (Wainer & Mislevy, 2000). Note also that if no prior information is available about the ability distribution, then $p(\theta)$ would simply be the uniform distribution (that is, a constant), often known as an *non-informative prior*. In such cases, the MAP estimator is the same as the MLE. In other words, MLE is a special case of MAP estimation with a non-informative prior (Parshall et al., 2002).

An alternative way to obtain the MAP ability estimate and its standard error is to use hierarchical Bayesian computational methods such as MCMC to sample from the posterior distribution. Statistical properties of the samples can then be used to draw inferences about the posterior distribution and to determine the MAP estimate and its precision (Wainer et al., 2007).

In contrast to MLE and MAP, EAP ability estimation can be performed non-iteratively. This is done by choosing a finite number of θ values along the ability scale called *quadrature nodes*, Q_r ($r = 1 \ldots q$, where q is the total number of nodes). Weights, denoted $W(Q_r)$, are assigned to each node. The weights are typically drawn from a known distribution such as the standard normal distribution and scaled so that they sum to 1.0 (Embretson & Reise, 2000). The likelihood function at each quadrature node, $L(Q_r)$ is evaluated and an EAP ability estimate is then derived using the formula below (Bock & Mislevy, 1982).

$$\hat{\theta} = \sum_{r=1}^{q} [Q_r \cdot L(Q_r) \cdot W(Q_r)] \Big/ \sum_{r=1}^{q} [L(Q_r) \cdot W(Q_r)] \tag{12}$$

Measurement precision of the EAP ability estimate is obtained by computing the standard deviation of the posterior distribution and is given as,

$$SE(\theta) = \sqrt{\sum_{r=1}^{q}[(Q_r - \theta)\cdot L(Q_r)\cdot W(Q_r)] \Big/ \sum_{r=1}^{q}[L(Q_r)\cdot W(Q_r)]} \qquad (13)$$

One of the advantages of EAP over MAP and MLE is that because it is non-iterative, it is computationally faster. This is important in the CAT context as provisional θ estimates typically need to be computed quickly so that the next item can be determined (Embretson & Reise, 2000). In addition, both Bayesian procedures are quite stable for short tests and are always bounded. Thus, it does not have the issue with MLE where θ cannot be estimated if examinees answer all items correctly or incorrectly. However, the Bayesian ability estimates do have some centrifugal bias – they tend to underestimate high abilities and over estimate low abilities (Parshall et al., 2002). This is also known as regression towards the mean and is an issue unless the number of items is large. Furthermore, the Bayesian procedures can be additionally biased if the prior distribution is not correctly specified (Wainer & Thissen, 1987).

Stopping Rule

Every CAT needs a stopping rule that determines when the item administration should terminate. CAT stopping rules fall generally into two categories resulting in two types of adaptive tests: fixed-length test and variable-length tests.

Fixed-Length CATs

Fixed-length CATs administer items until a predetermined number of items have been given. Thus, each examinee receives the same number of item on the test. Fixed-length CATs have the advantage of being easier to implement and having better prediction of item pool usage rates (Thissen & Mislevy, 2000). Also, fixed-length tests

give the perception of fairness and are easier to explain to examinees. If different examinees receive different number of items, then examinees who perform poorly with relatively short tests may claim that they did not get the same opportunity as others to prove their competence (Bergstrom & Lunz, 1999). As such, fixed-length CATs are very popular and have been implemented for CATs in a variety of assessments. Examples include the CAT version of the Armed Service Vocational Aptitude Battery (CAT-ASVAB; Segall & Moreno, 1999), the CAT version of the GRE (Mills, 1999), and the certification exam of the American Society of Clinical Pathology (ASCP; Bergstrom & Lunz, 1999). Administering fixed-length CATs, however, negates one of the main benefits of an adaptive test over a non-adaptive one. That is, adaptive tests can measure examinees across the ability scale to the same level of precision. This can only be accomplished when the test length is allowed to vary (Thissen & Mislevy, 2000).

Variable-Length CATs

A variable-length CAT tests each examinee until a pre-specified level of measurement precision is reached. The criterion for stopping can be a target standard error (SE) of measurement for MI selection or a target posterior precision under Bayesian selection (Thissen & Mislevy, 2000). For a criterion-reference test (CRT), it can also be a specified level of confidence in the pass/fail decision (Bergstrom & Lunz, 1992; Kingsbury & Weiss, 1983). The main advantage of variable-length CATs is that every examinee is measured to the same degree of precision. Examinees with ability well-targeted by the items in the pool (for example, around the mode of the ability distribution) generally receive shorter tests than those with ability levels in the extremes (Parshall et al., 2002). This generally results in more optimal use of the item pool since many examinees take a minimal-length test, and more efficient use of the examinees' time and effort (Bergstrom & Lunz, 1999; Segall & Moreno, 1999). However, it may be

difficult to explain to examinees and their constituents why comparable decision or evaluations can be made based on tests of different lengths. In addition, tests may need to be lengthened and made equal to ensure equivalent content coverage (Bergstrom & Lunz, 1999). Examples of variable-length CATs include several national licensure and certification tests such as the National Certification Examination (NCE) for registered nurse anesthetists and the National Council Licensure Examination of Registered Nurses (Bergstrom & Lunz, 1999). It is also possible to use a combination of the types of stopping rules. Thissen & Mislevy (2000), for example, advise that some mixture of "target precision" and "maximum number of items" should always be used in practice so that some specific measurement precision can generally be achieved unless the item pool runs out of appropriate items to administer.

The four components of CAT described above relate directly to the measurement precision and efficiency of an adaptive test. Two additional issues – those of satisfying content-related test specifications and ensuring test security – typically need to be addressed for a CAT to be acceptable for its specific testing purpose. Addressing these issues invariably leads to a trade-off between measurement efficiency and the acceptance of the CAT as a viable and defensible mode of testing (Parshall, Davey & Nering, 1998; Stocking & Lewis, 2000). They are resolved by constraining the testing algorithm with *content balancing* and *exposure control* procedures.

Content Balancing

Tests are usually constructed to satisfy a set of rules known as a test specification or test blueprint. With a traditional P&P test, test constructors pre-assemble the test forms to conform to the test blueprint. With a CAT, however, the test specification must be implemented as part of the item selection algorithm (Kingsbury & Zara, 1991). Stocking and Swanson (1993) classified the types of test specification constraints into

four categories: constraint based on intrinsic item properties (such as item content and format), overlap constraints (such as cross-information and redundancy across items), item set constraints (such as items with a shared stimulus), and constraints of statistical properties (that is, psychometric constraints). Procedures that ensure the satisfaction of the first type of constraint are known as *content balancing* procedures. Two popular content balancing procedures are the constrained CAT procedure (Kingsbury & Zara, 1989) and the weighted deviations model (Swanson & Stocking, 1993).

Constrained CAT Procedure

Kingsbury and Zara (1989) propose the use of a constrained CAT (C-CAT) procedure within the maximum information selection algorithm. This is done by pre-specifying, according to the test blueprint, the percentage of items that need to be administered in each content area or of each item format. For example, a math test may require 20% addition, 20% subtraction, 30% multiplication and 30% division items. These percentages become *sub-goals* within each content area or item format. As items are administered to each examinee, the proportion of items given under each sub-goal is tracked. The next item to administer is determined by finding the sub-goal with the largest discrepancy, then looking within the sub-goal for the item that provides the most information at the provisional ability estimate for the examinee. The first item administered can be from a randomly chosen sub-goals or from the sub-goal with the largest percentage requirement (Boyd, 2003).

The C-CAT procedure has the advantage of being relatively simple to understand and implement. It has been used in certification and licensure CATs (Bergstrom & Lunz, 1999; Zara, 1989) and in adaptive tests within the K-12 educational setting (Kingsbury, 1990). The major disadvantage of the C-CAT procedure is the assumption that it is feasible to partition the item pool into mutually exclusive subsets based on the content

areas, item formats or other item features that require balancing. As the number of content constraints increases, the number of mutually exclusive partition grows at a high rate, and the number of available items in each partition becomes quite small, making the testing algorithm less efficient (Stocking & Swanson, 1993). A more sophisticated content balancing procedure is needed to overcome this issue.

Weighted Deviations Model

The weighted deviations model (WDM) is a comprehensive methodology proposed by Swanson and Stocking (1993) not only to handle constraints based on intrinsic item properties, but also overlap, item set, and statistical constraints. It is derived from the *binary programming* model. In this model, test constraints are mathematically formulated as linear equations or inequalities. An objective function based on targeted test information functions or posterior variances is then optimized (maximized or minimized) subject to the test constraints. The binary programming model has been applied generally in the area of automated test assembly (ATA; van der Linden, 1998). WDM enhances the traditional binary programming model by treating test constraints as desired properties, assigning each constraint weights depending on its relative importance, and moving the constraints into the objective function (Stocking & Swanson, 1993). Then, at each step of the adaptive test, the item that optimizes the objective function is chosen for administration.

WDM has the advantage of being able to handle a large set of constraints. It has been shown to have satisfactory performance when applied to real item pools (Stocking, Swanson & Pearlman, 1991, 1993). However, as it is a mathematically sophisticated procedure, many practitioners find it difficult to understand and too complicated to implement, and often choose a simpler strategy such as the C-CAT procedure.

Exposure Control

An unconstrained testing algorithm that chooses the best items, based on either MI or Bayesian selection, often selects similar sets of items for examinees at start of the test, especially if no prior information about each examinee is known (Thissen & Mislevy, 2000). Examinees with similar abilities can also end up with significant overlap in the sets of items they are given over the entire test. Several studies have found that for unconstrained CAT algorithms, certain items are administered to nearly every examinee and a small portion of items in the pool account for a large proportion of the item administrations (Hulin, Drasgow & Parsons, 1983; Mills & Stocking, 1995; Parshall et al, 2002). This unbalance in pool utilization is undesirable economically because it is a waste of item development efforts to construct a large item pool in which a substantial portion of items are seldom or never administered (Revuelta & Ponsoda, 1998).

The overexposure of often-administered items also leads to test security concerns. One major benefit of CAT over traditional P&P tests is the flexibility in examinee access to the test, allowing examinees to take the test at various locations and at different times (Wainer & Eignor, 2000). However, this flexibility can also lead to a compromise in test security and a threat to test validity if examinees share information with one another between test administrations. Kaplan Educational Centers demonstrated the severity of this issue when, in 1994, it sent employees to take the CAT-GRE and memorize test items. Within a short time period, most of the items being reported back were already on its list of compromised items. The administration of the GRE was temporarily suspended after Kaplan informed ETS of its findings (Davey & Nering, 2002). Since then, further incidents involving the widespread sharing of test items leading to breaches in test security have forced ETS to shut down its CAT-GRE and bring back P&P testing (Chang, 2004). Such test security concerns highlight not only the importance of having a

large and frequently replenished item pool (Way, 1998), but also the need for the testing algorithm to control the exposure of items in the pool.

As such, an exposure control procedure should have two main goals: to prevent the overexposure of often-administered items and to increase the usage rate of seldom- or never-selected items (Revuelta & Ponsoda, 1998). Various exposure control strategies have been proposed in recent literature and these approaches can be classified into three general categories: randomization procedures, conditional procedures and stratification procedures (Way, 1998; Davis & Dodd, 2003; Davis, 2004).

Randomization Procedures

Randomization procedures control the selection of items by randomly choosing the next item from a set of near-optimal items instead of always choosing the most informative one. The various randomization strategies differ in how the near-optimal sets of items are formed and the sizes of these sets. Examples of randomization exposure control procedures include the 5-4-3-2-1 technique (McBride & Martin, 1983; Hetter & Sympson, 1997), the randomesque method (Kingsbury & Zara, 1989), the "choose one of three" randomization procedure (Thomasson, 1998), the within 0.10 logits method (Lunz & Stahl, 1998), and the progressive method (Revuelta & Ponsoda, 1996). The advantage of randomization procedures is that they are easy to understand and relatively straightforward to implement. However, they provide no guarantee that exposure rates of items will be constrained to a given level (Davis, 2004).

Conditional Procedures

Conditional strategies control the probability of each item being administered at each step in the test, conditional on a given criterion. The criterion is typically based on exposure control parameters, which limit the maximum exposure rate of each item to a

predetermined level. Thus, conditional procedures have the advantage of providing a guaranteed maximum exposure rate. However, they often require complex time-consuming simulations to determine the exposure control parameters. These simulations need to be performed prior to the operational use of the CAT and can increase implementation complexity (Davis & Dodd, 2003). Examples of conditional exposure control strategies include the Sympson-Hetter procedure (Sympson & Hetter, 1985), the conditional Sympson-Hetter procedure (Stocking & Lewis, 1998), the Davey-Parshall procedure (Davey & Parshall, 1995), the Stocking and Lewis multinomial procedure (Stocking & Lewis, 1995), the restrictive maximum information method (Revuelta & Ponsoda, 1996), and the tri-conditional procedure (Parshall, Hogarty & Kromrey, 1999).

In addition, Revuelta and Ponsoda (1998) proposed a combination of the progressive procedure and the restrictive maximum information procedure known as the *progressive-restrictive* procedure. It is hence a hybrid approach with both randomization and conditional components. Because the progressive-restrictive procedure is implemented in this dissertation, further details about this approach will be given in a later section.

Stratification Procedures

With stratification procedures, the item pool is partitioned into strata based on every item's discrimination (a) parameter. The resulting strata are arranged from low to high discriminating power. The test is then divided into stages that match the number of strata and a predefined number of items are administered from each stratum in the corresponding stage of the test. The rationale behind this stratification strategy is to administer the least discriminating items at start of the test where the accuracy of the ability estimation is low, saving the highly peaked informative items for the latter stages of the test when the examinee's ability needs to be pinpointed (Chang & Ying, 1996;

Davey & Nering, 2002). Forcing the items with low *a*-value to be administered at the beginning of the test would also more evenly distribute the utilization of items across *a*-values and hence control item exposure within the entire item pool (Chang, 2004). Examples of stratification procedures include the *a*-stratified design (Chang & Ying, 1996), the *a*-stratified design with *b*-blocking (Chang, Qian & Ying, 1999), a multiple-stratification variant of the *a*-stratified design that incorporate content balancing (Yi & Chang, 2000), and the enhanced *a*-stratified design (Leung, Chang & Hau, 1999).

Exposure Control with Testlets

The exposure control procedures described above were proposed for CATs with independent and dichotomously-scored items. While these methods work well for dichotomous items, the results may not generalize to CATs with polytomous items or CATs containing testlets.

CATs with polytomously-scored items typically have smaller item pools and the mode of the information curve for such items is usually more spread across the ability scale than dichotomously-scored item pools. Thus, the negative effects on measurement precision from administering suboptimal items may not be as pronounced as for CATs with dichotomous items (Koch & Dodd, 1989). Research in exposure control for CATs with polytomous items is not as extensive and most studies are quite recent (e.g. Pastor, Dodd & Chang, 2002; Davis, Pastor, Dodd, Chiang & Fitzpatrick, 2003; Davis & Dodd, 2003; Boyd, 2003; Davis, 2004). Furthermore, the results from these studies only apply to testlet-based CATs if the testlets are scored using one of the polytomous IRT models.

Research in exposure control procedures for testlet-based CATs measured with TRT is even more limited. Boyd (2003) compared several exposure control procedures in testlet-based CATs measured with the 3PL-TRT. The study included the randomesque method, the Sympson-Hetter procedure, the progressive-restrictive procedure, and a

modification of the within 0.10 method to accommodate polytomous items and testlets (Davis & Dodd, 2003). The study found that the Sympson-Hetter procedure, a conditional procedure, was able to maintain the maximum exposure rate, it had very poor pool utilization, with about 60% of the testlet in the pool never administered. The randomization procedures (i.e. the randomesque and modified within 0.10 methods) yielded low maximum exposure rates, but still had about 30% of the pool not utilized. The progressive-restrictive procedure yielded the optimal results in that it controlled for the maximum exposure rate while consistently utilizing all testlets in the pool. It should be noted, however, that the level of selection for the testing algorithms in Boyd's (2003) study was at the *testlet* level and not at the item level. In other words, instead of choosing the next item to administer, the testing algorithm determined the next *testlet* to administer based on the set of items pre-assigned to each testlet. No studies to date have examined exposure control procedures for testlet-based CATs whose testing algorithm's level of selection is at the item level.

The Progressive-Restrictive Procedure

Revuelta and Ponsoda (1998) devised the progressive-restrictive procedure as a method that leverage the advantages both a randomization procedure (the progressive method) and a conditional procedure (the restricted maximum information method).

The progressive method (Revuelta and Ponsoda, 1996) extends the unconstrained MI item selection by adding a random component that affects the chances of an item being chosen. Suppose that the serial position of the current item is defined as $s = h/m$, where h is the number of items administered so far and m is the maximum length of the test. Suppose also that I_i is the information provided by an unused pool item (i) conditional on the provisional ability estimate. The random component (R_i) for that unused item is then a value randomly drawn from the uniform (0, H) distribution, where

H is the maximum I_i value among all the unused pool items. A weight (w_i) is computed for each unused item as a linear combination of the random and information components according to the formula,

$$w_i = (1 - s) \cdot R_i + s \cdot I_i \qquad (14)$$

The item with the highest w_i value at each point in the test is chosen for administration. The rationale behind this method is that at the beginning of a test, when less is known about the examinee, the random component contributes more to the chances of an unused item being chosen. As the test continues and more is learned about the examinee, the information component contributes progressively more to the probability of an unused item being chosen while the contribution of the random component diminishes. As with most randomization procedures, the progressive method shows adequate measurement precision while significantly improving pool utilization – all items in the pool are generally administered at once. However, it provides little control over the maximum exposure rate such that some items – those with high discrimination parameters – are administered far more frequently than others (Revuelta and Ponsoda, 1998).

The restrictive maximum information method (Revuelta and Ponsoda, 1996) was proposed as an alternative to the complex Sympson-Hetter procedure (Sympson & Hetter, 1985). An exposure control parameter, k, is specified such that no item in the pool is exposed in more than $100k\%$ of the examinees. Suppose that the CAT has been administered to t examinees so far, and an item (i) has been given a_i times. The exposure rate r_i is, therefore, a_i/t. For the next examinee, the only items eligible to be administered are those with $r_i < k$ and those items are chosen based on MI selection. Thus, items that are administered frequently at first would quickly become ineligible for administration, allowing other items the chance to be selected. As more examinees are tested, the exposure rates for ineligible items decrease and the item would eventually become

eligible for administration again. The restrictive maximum information method, like other conditional procedures, provides acceptance measurement precision and keeps the maximum exposure rate under that specified by the control parameter. However, it shows poor pool utilization in that a substantial number of items in pool are never administered.

The progressive-restrictive procedure includes aspects of the two aforementioned methods. As in the restrictive maximum information method, it does not allow any item to be administered to more than 100*k*% of the examinees. However, instead of using MI selection, it uses formula (14) in the progressive method to choose the next item from the set of eligible items in the pool. The progressive-restrictive procedure has been shown to produce good measurement precision when the exposure control parameter, k, is sufficiently high (for example, $k = .40$). It also has consistently utilizes all items in the pool while keeping the maximum exposure rate under control (Revuelta and Ponsoda, 1998). Recall that Boyd (2003) also found that the progressive-restrictive procedure was the optimal exposure control method for testlet-based CAT measured with TRT as well as with a polytomous IRT model. As such, it will be the exposure control procedure used in the CAT conditions for this dissertation.

MULTISTAGE TESTING

Multistage testing is a form adaptive testing that has become increasingly popular, especially as an alternative to CAT in computer-based testing. The idea of multistage testing is not a new one. Cronbach & Glaser (1965), Lord (1971, 1974), Weiss (1973) and Loyd (1984) each proposed P&P versions of multistage test designs. Research into these designs, however, was eclipsed by the development and implementation of CAT in literature and in practice (Mead, 2006). In recent years, however, the adaptation of multistage testing into a computerized testing framework coupled with concerns about

the practical shortcomings of CAT has renewed interest in this alternative form of adaptive testing.

If one were to put the degree of test adaptation on a continuum, then traditional P&P tests would be at one extreme, where the tests are linear with no point of adaptation, resulting in the same set of items given to all examinees taking the same test form. On the other extreme would be item-level CATs, where adaptation occurs after every item and each examinee could potentially receive a completely different set of items. Multistage tests (MSTs) would then be considered a middle ground between these two extremes (Jodoin, Zenisky & Hambleton, 2006). A MST does not have its points of adaptation at the item level, but is instead adaptive between sets of items, called *modules*. Examinees are typically administered a common initial module. Then, based on their performance, they are adaptively routed to different sets of items in the later parts of the test. Consequently, MSTs generally have improved measurement precision and efficiency over tradition P&P tests (Luecht, Nungester & Hadadi, 1996). More similar to the P&P tests, however, the modules that can be given and the routes that can be taken through a MST are constructed *prior* to test administration. This allows MSTs to overcome some of the common criticisms of item-level CATs.

One such criticism is the lack of administrative control over the content quality of CAT test forms. Every CAT test form is built on-the-fly during test administration and hence cannot be reviewed apriori by test developers, particularly content experts. Even with the most sophisticated content balancing algorithm, it is generally not possible to code every content specification in test blueprint. Also, context effects due to item ordering or other factors, such as one item cuing the answer to another, can be difficulty to account for in testing algorithms. Thus, the use of human review for quality assurance is often essential (Luecht & Nungester, 1998; Patsula, 1999). MSTs usually use testing

automated test assembly (ATA) algorithms to build the initial item sets such that they satisfy the various statistical and content constraints. Test developers are then given the opportunity to review and edit the item sets for content quality. Item exposure can also be controlled through this quality assurance process prior to test administration (Patsula, 1999; Hendrickson, 2007).

Another criticism of item-level CATs is the lack of review opportunities for examinees. To prevent examinees from using test-taking strategies that would circumvent the testing algorithm and threaten the measurement efficiency of the test, most CATs prohibit examinees from reviewing and skipping items during the test (Lunz & Bergstrom, 1994; Vispoel, 1998). However, examinees find this to be a big disadvantage and it is one of the greatest complaints about CAT from test takers (Patsula, 1999). MST can overcome this issue by allowing examinees to review and edit their item responses after each item set without concerns for the integrity of the test.

As such, the development and implementation of multistage testing have increased substantially in recent years. MSTs has been given various names and appeared in different formats, including computer mastery testing (CMT; Lewis & Sheehan, 1990), computer-adaptive sequential testing (CAST; Luecht & Nungester, 1998), multiple form structures (MFS; Armstrong, Jones, Koppel & Pashley, 2004) and bundled multistage adaptive testing (BMAT; Luecht, 2003). Several large-scale assessments have also been implemented as MSTs. Examples include the Law School Admissions Test (LSAT), the Test of English as a Foreign Language (TOEFL), the National Council of Architectural Registry Board (NCARB), the National Assessment of Educational Progress (NAEP), the U.S. Medical Licensure Examination (USMLE) and the Uniform CPA (certified public accountant) Examination (Luecht, Brumfield & Breithaupt, 2006; Hendrickson, 2007).

Components of Multistage Tests

Under the broadest definition, an item-level CAT can be considered a special case of a MST whose point of adaptation is after each item. However, it is commonly understood that an MST is an adaptive test that adapts after more than just one item and whose item sets are pre-assembled (Hendrickson, 2007). Because of the similarities though, the basic components of an MST are not too different from those of a CAT. Every MST should consist of an item pool from which the test is built, a testing algorithm (or routing method) that controls how items or sets of items are administered, and an ability estimation method that computes provisional and final ability estimates. Because the length of a MST is typically pre-specified, it has a fixed-length stopping rule. Additionally, a MST has several unique components. Using the terminology developed by Luecht and Nungester (1998), these components include modules, panels, and stages.

As defined earlier, *modules* are blocks of items that are built before they are administered to examinees. They are also often referred to as *item blocks* or *testlets* (Jodoin et al., 2006). However, to avoid confusion in this dissertation, the term testlet will be reserved for a set of items from a single content area that are related through a common stimulus, such as a reading passage, diagram or graph (Wainer & Kiely, 1987). Thus, modules tend to be larger units than testlets; a single module may consists of more than one testlet. Each individual module in a MST is often built to satisfy statistical targets such as test information functions (TIF) while being comparable to other modules in content coverage. Modules may hence be classified in terms of their overall difficulties as, for example, easy, moderate and hard modules (Luecht and Nungester, 1998).

Once modules are built, they are grouped into test administration units called *panels*. Each panel is a particular combination of modules that satisfy content and statistical constraints specified in the test blueprint. Panels are therefore analogous to test

forms on a traditional P&P test. To control for the exposure of modules and items, a number of panels are usually assembled for a MST administration and each examinee is randomly administered one of the panels.

Within a panel, modules are arranged into a series of *stages*. MSTs can have any number of stages, but most typically have two or three stages. Within each stage, there could also be a number of modules. The first stage of a MST, however, is usually a single module taken by all examinees who are assigned the particular panel. The later stages can have any number of modules, although three or four modules in each stage have been shown to be sufficient (Armstrong et al., 2004; Hambleton & Xing, 2006).

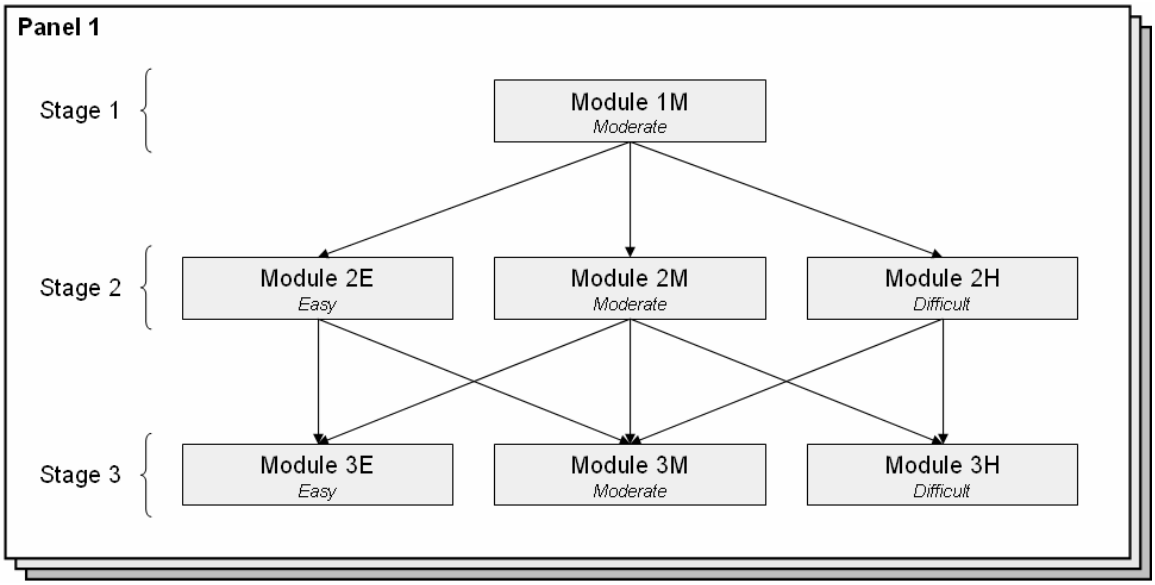

Figure 2.2: Example Multistage Test with 1-3-3 Stage Structure

Figure 2.2 above gives an example of one of the panels in a seven-module, three-stage MST. This MST is termed a *1-3-3* stage structure design because it has 1 module in the first stage and 3 modules each in the second and third stage within each panel (Luecht, 2000). The 1-3-3 MST design has been implemented in several studies (e.g.

Luecht & Nungester, 1998; Davis & Dodd, 2003; Luecht et al., 2006; Jodoin et al, 2006; Hambleton & Xing, 2006; Chuah, Drasgow & Luecht, 2006).

Any examinee who is administered the panel in Figure 2.2 would take the set of items within Module 1M of Stage 1, which is of moderate difficulty. Then, depending on the examinee's performance on this first module, one of the three Stage 2 modules is administered. Examinees that perform poorly in Stage 1 are routed to the easy Stage 2 module (Module 2E); examinees with moderate performance are routed to the moderate Stage 2 module (Module 2M); and examinees that performed well in Stage 1 are given the items in the difficult Stage 2 module (Module 2H). Similar rules are used to route examinees from the Stage 2 modules to the Stage 3 modules. The criteria for determining which module an examinee should be routed to in the next stage is implemented in the testing algorithm or routing method. Specifying the details of the routing method is an example of the several MST design decisions that need to be considered when implementing a MST.

Multistage Test Design Considerations

Design decisions about the various components of a MST need to be considered and made before its implementation and operational use. These include decisions about the item pool, test structure, routing method, scoring and ability estimation, and test assembly. Factors such as the purpose of the test, the ability distribution of the examinees, the type of test items and the desired content coverage of the test all need to be considered when making these decisions. Details of these test design considerations as well as studies that have examined them are summarized in the sections to follow.

Item Pool

Many of the item pool considerations for a CAT also apply to item pools for a MST. The items in the pool need to be developed to satisfy the content and psychometric requirements of the test. The size of the pool should be sufficiently large to support the assembly of modules within multiple panels (Hendrickson, 2007). And just like for a CAT, the purpose of a MST (norm-referenced vs. criterion-referenced) also affects how statistical information in the item pool is distributed (Urry, 1977; Reckase, 1981; Parshall et al., 2002).

Two recent studies have examined the impact of item pool characteristics on the psychometric properties of MSTs and compared it with that of computer-based linear fixed length tests (LFTs) and CATs (Xing, 2001; Jodoin, 2003). Xing's (2001) study compared the impact of several test design variables on the measurement precision of tests whose primary purpose is to make pass fail decisions. The variables investigated included item pool size, item pool quality and the three computer-based test designs. Across all test designs, the study found that improvements in item pool quality, measured by the discrimination parameters, increased decision accuracy and consistency of the test. Doubling item pool size (from 240 to 480 items) had little impact of the measurement properties, but helped decrease item exposure significantly. Jodoin (2003) compared the three computer-based test designs on several additional psychometric measures. The item pool characteristics manipulated in the study included item pool quality and the match between test and item pool content specifications. For all three designs, the study found that test reliability, overall and conditional measurement precision and classification precision increased with item pool quality, and decreased when the match between item pool and test specification was less adequate.

Test Structure

Numerous design decisions need to be made about the structure of a MST. They include issues such as the number of stages, the number of modules within each stage, the number of items or testlets within each module and the overall test length. The decisions for these issues are often dictated by the specific requirements imposed by the test blueprint or available items. However, in most cases, test developers have options to consider and several studies have evaluated these options.

Number of Stages. Much of the earlier research on MST used only two stages (e.g. Cronbach & Glaser, 1965; Lord, 1971; Loyd, 1984; Kim & Plake, 1993). However, this was done mainly in the context of making a P&P test adaptive. With the administrative expediency afforded by computerized testing, most recent research and applications used three or four stages (Hendrickson, 2007). Pastula (1999) found that increasing the number of stages from two to three increased the accuracy in ability estimation. Jodoin et al. (2006), however, found that while a two-stage 40-item MST performed worse then a three-stage 60-item MST in terms of ability estimation and classification precision, the difference was slight and the results for the two-stage MST were generally acceptable from a practical perspective. These results highlight the point that while increasing the number of stages generally increases the measurement precision of a MST, the psychometric gains needs to be balanced against the increase in complexity of test construction (Luecht et al., 1996; Luecht & Nungester, 1998).

Number of Modules Within Each Stage. The vast majority of MST research and applications have used one module in the first stage and multiple modules in the later stages. For example, as noted earlier, the 1-3-3 stage structure design (in Figure 2.2) is a common implementation in MST research. Early studies showed that the number of modules in the second stage had an impact on measurement accuracy (Lord, 1971; Kim

& Plake, 1993). Patsula (1999) also found that increasing the number of modules from three to five in later stages also increased the accuracy in ability estimation. Zenisky (2004), however, in comparing four MST across-stage module arrangements (1-2-2, 1-3-3, 1-2-3, and 1-3-2) found no differences in decision accuracy and consistency across the four arrangements. Thus, as with the number of stages, the choice in the number of modules per stage is also a decision that requires tradeoff between measurement precision and test assembly complexity. Armstrong et al. (2004) found evidence indicating that three modules per stage is generally adequate while having four modules is the maximum needed to achieve desirable psychometric properties for most MSTs.

Number of Items or Testlets Within Each Module. The number of items within each module ranges from 1 to 90 in research studies and operational MSTs, with a mean of about five items per module (Hendrickson, 2007). Two early studies (Loyd, 1984; Kim & Plake, 1993) investigated the impact that the length of the first-stage module (also known as the *routing test*) had on ability estimation and found that longer routing tests were better. However, the total test length in neither of these studies was fixed. Patsula (1999) was one of the first studies to systematically examine the effect of varying the number of items within each module in a fixed-length test. The study found that, at most ability levels, varying the number of items per module had little impact on the accuracy and precision of ability estimation. Davis and Dodd (2003), in comparing the measurement precision and exposure control of testlet-based CATs and MSTs, employed a MST design with three testlets in the first-stage module, and one testlet each in the second- and third-stage modules. The study found that this design yielded superior overall performance compared to other CAT methods. No studies in the literature investigated, however, has systematically examined the effect of the number of testlets within each module.

Overall Test Length. It is well-known from the Spearman-Brown prophecy formula (Crocker & Algina, 1986) that increasing test length by adding items of comparable psychometric quality increases the reliability of the test, and therefore, improves measurement precision. This was also demonstrated empirically in early MST research (Loyd, 1984; Kim & Plake, 1993). Thus, few studies have systematically investigated the impact that varying the overall test length. Jodoin (2003) and Jodoin et al. (2006) compared 40-item and 60-item MSTs and in both cases, the psychometric gains from increasing the test length by 50% were only modest. Stark & Chernyshenko (2006), however, noted that the test lengths in Jodoin et al. (2006) as well as in other MST studies were likely too long to reveal any psychometric benefits of different test lengths, especially in comparison to traditional P&P tests and LFTs. They postulate that the greatest impact on ability estimation and classification precision for MSTs may simply lie in the overall test length instead of the other test design factors and suggest that future research investigate this claim by including much shorter test.

Routing Method

The routing method for an MST is analogous to the test algorithm in a CAT. It determines, based on the performance of each examinee in the previous stage, which module the examinee should be routed to in the next stage. Lord (1980) notes that the routing method is a particularly critical element to the usefulness of an adaptive multistage test. As such, several studies have examined and compared strategies for routing examinees in a MST. Two routing strategies mentioned in recent MST studies are the defined population interval and approximate maximum information methods.

The goal of the defined population interval (DPI) method is to route examinees so that specific proportions of the examinee population would be expected to take the various modules in the next stage (Luecht et al, 2006). This involves pre-defining the

expected proportions of examinees for each route, then determining the routing cut points that would lead to the desired proportions. The cut points are typically first determined on the ability (θ) scale and then transformed into the number-correct scores. For example, if, in a 1-3-3 MST stage structure, we would like one-third of the examinees to be routed to each of the three stage-2 modules and to each of the three stage-3 modules, then ability (θ) values associated with the 33rd and 67th percentiles of the cumulative distribution of θ should be determined. If θ is assumed to be normally distributed, then $\theta_1 = -0.44$ and $\theta_2 = 0.44$ would be the two cut points. These θ cut points are then transformed to number-correct cut points by computing,

$$X_1 = \sum_{i \in \text{current stage}} P(\theta_1; \xi_i) \text{ and } X_2 = \sum_{i \in \text{current stage}} P(\theta_2; \xi_i) \quad (15)$$

Where $P(\theta; \xi_i)$ represents the item response function for the chosen measurement model (e.g. 1PL-, 2PL-, 3PL-IRT or TRT) and ξ_i is the set of item parameters associated with the measurement model. Examinees are then routed to either the easy, moderate or difficult module in the next stage based on their number-correct score at the current stage relative to X_1 and X_2 (Luecht et al, 2006). The DPI method is relatively simple to implement and has been widely used in recent MST studies (Xing, 2001; Jodoin, 2003; Zenisky, 2004; Jodoin et al., 2006; Hambleton & Xing, 2006; Chuah et al., 2006). It does, however, require making distributional assumptions about θ. It may also not be appropriate for criterion-referenced tests as it is fundamentally a norm-referenced methodology (Zenisky, 2004).

The approximate maximum information (AMI) method determines the routing cut points based on the cumulative test information functions (TIFs) of modules within the same stage. It finds the θ cut points by finding the intersections of the TIFs for adjacent modules within a stage. So, for example, to find the cut points for routing between Stage

1 and Stage 2 of a 1-3-3 MST (see Figure 2.2), θ values corresponding to the intersection of the TIF curves for (1M+2E) and (1M+2M) as well as (1M+2M) and (1M+2H) are computed using standard numerical analysis root-finding techniques (Luecht et al, 2006). The θ cut points are then transformed to number-correct scores using the same formulas in Equation (15). The advantage of the AMI method is that it determines the routing cut points empirically using a maximum information criterion similar to that in CAT. Thus, it is expected to have good measurement precision. However, the process for determining the θ cut points usually needs to be repeated for each panel, unless care is put into making the TIFs across multiple panels virtually identical (Luecht et al, 2006).

Many additional MST routing strategies have been proposed in literature. Zenisky (2004) provides an excellent review of the various routing methods. Because the effectiveness and properties of the numerous routing methods are still not fully understood, this is an area of MST where further, more systematic research is needed (Stark & Chernyshenko, 2006).

Scoring and Ability Estimation

Closely related to decision of the routing method is the choice of how to score the modules and test and estimate examinee ability. As seen in the previous section, number-correct scoring transformed from the θ scale can be used to route examinees from one stage to the next. Doing so, however, requires a choice in the underlying measurement model for the θ scale. Research and applications of MSTs have often used the 3PL-IRT model for dichotomously-scored items or the nominal or graded response model for polytomous items (Hendrickson, 2007). MSTs with modules that consist of multiple testlets have been scored using the partial credit model (Davis & Dodd, 2003). The use of TRT for scoring modules (with or without testlets) has been suggested in recent literature (Zenisky, 2004; Hendrickson, 2007) because it would capture any dependency

between items not only within a testlet but also within a module. However, no study in the literature investigated has implemented TRT in the context of MST.

While number-correct scoring is a straightforward method for routing examinees between MST stages, it would not be appropriate to use number-correct scoring on the overall test as the final ability estimate for each examinee. This is because, like in a CAT, examinees of a MST do not received statistically equivalent items (Lord, 1980). The same methods for estimating ability in CATs are also applicable to MSTs and examples of using MLE (e.g. Kim & Plake, 1993; Davis & Dodd, 2003; Jodoin et al., 2006; Chuah et al., 2006), MAP (e.g. Schnipke & Reese, 1997) and EAP (e.g. Jodoin, 2003; Luecht et al., 2006; Hambleton & Xing, 2006) have been found in MST literature.

Test Assembly

Once decisions about the MST test design considerations above have been made, a method of assembling the modules and panels also needs to be chosen. Test assembly for MST is a very complicated process because it involves simultaneously generating multiple panels that are parallel in both content and psychometric properties. These panels must consist of modules that meet specific statistical requirements such as target TIFs (Luecht & Nungester, 1998). In addition, constraints related to content balancing, exposure control, context effects, examinee cognitive levels, item and testlet overlap, item format, word count and other characteristics of interest or concern also need to be satisfied (Hendrickson, 2007). Consequently, automated test assembly (ATA) algorithms and computer programs are often utilized to solve the test assembly problem in MSTs.

The development of ATA algorithms actually preceded that of MST and occurred in the more general context of optimal test design and assembly (Birmbaum, 1968; van der Linden, 2005); particularly for the mass construction of computer-based tests (Parshall et al., 2002). van der Linden (1998) provides an excellent summary of the

general approaches to ATA and the rich amount of studies under each of these approaches. Several studies have tailored ATA algorithms to the assembly of modules and panels in MSTs (e.g. Adema, 1990; Luecht & Nungester, 1998; Breithaupt, Ariel & Veldkamp, 2005; Ariel, Veldkamp & Breithaupt, 2006; Luecht et al., 2006). MST-specific ATA computer programs, such as CASTISEL (Luecht, 1998), have also been written and are widely available for assembling MSTs. If the constraints for a MST are relatively few and manageable, then manual test assembly without the aid of ATA software is also possible (e.g. Davis & Dodd, 2003).

In general, panels in for a MST can be assembled using one of two strategies: bottom-up or top-down (Luecht & Nungester, 1998). With the bottom-up strategy, each module is built to *module-level* specifications for statistical targets, content and other test features. That is, each module is like an independent mini-test and modules built to the same module-level specifications are exchangeable across panels. In contrast, with the top-down strategy, modules are constructed according to *test-level* specifications. Thus, modules are dependent of one another and must be combined to satisfy the test-level requirements; they are not exchangeable across panels. Luecht & Nungester (1998) provides examples of panels assembled using each of the two strategies. Examples of bottom-up (e.g. Luecht et al., 2006) and top-down (e.g. Davis & Dodd, 2003) strategies can also be found in MST literature.

Comparisons of MST with CAT

Because multistage testing is typically characterized as an alterative computer-based test design to the item-level CAT, numerous recent studies have compared the performance of the two designs. In this section, an overview of these comparability studies is provided.

Kim and Plake (1993) was one of the first studies to compare the measurement properties of MSTs and CATs. They compared the accuracy and relative efficiency in ability estimation of 18 simulated two-stage MSTs to three fixed-length CATs. The simulated MSTs varied in the length of the first-stage module (10, 15 or 20 items), distribution of item difficulty in the first-stage module (peaked or rectangular), and number of second-stage modules (6, 7 or 8 modules). Each of the second-stage modules had 30 items, resulting in MSTs with total test lengths of 40, 45, or 50 items. The fixed-length CATs differed in their total test lengths (40, 45 or 50 item) so that they were comparable to the MSTs. The item pools used by both test designs contained 354 simulated items measured with the 1PL-IRT model. MLE was used in both designs to estimate the ability of 1,600 simulated examinees. The study found that the fixed-length CAT outperformed two-stage MST of equal length in both accuracy and relative efficiency of ability estimation. Within the two-stage MST conditions, those that had a stage-one module with rectangular item difficulty distribution and an odd number of stage-two modules resulted in the most accurate ability estimates.

Luecht, Nungester and Hadadi (1996) were interested in the effect that various content balancing strategies had on the measurement properties and item exposure of CATs and MSTs. Three CATs with different content balancing methods were compared to two MSTs with different goals. One MST had a 1-2-3-4 stage structure and was designed to maximize the accuracy of ability estimation for most examines; that is, it was ideal of norm-referenced tests. The other MST had a 1-3-5 stage structure and its goal was to minimize mastery (pass-fail) decision errors; in other words, it was optimized for criterion-referenced tests. All CAT and MST conditions had an overall test length of 180 items, were constructed from an item pool consisting of 2,538 items whose parameters were based on real data and calibrated with the 1PL-IRT model. The

empirical ability estimates and item responses of 20,000 examinees who had previously taken these real items were used. The two different ability estimation procedures (MLE and EAP) were used for each of the five test conditions. The study found that all five conditions were generally similar and accurate in terms of ability estimation and mastery decisions. The two MST conditions were less efficient than two of the CAT conditions, but the difference in efficiency was acceptable given the administrative advantages of MSTs. Relatively little difference were found between the two ability estimation procedures across the test conditions. Lastly, the item exposure rates of the CAT conditions were higher than that of the MST conditions. However, it should be noted that no exposure control procedure was implemented with the CAT conditions.

Schnipke and Reese (1997) compared the psychometric properties of several testlet-based adaptive test designs. The designs compared included two variants of a two-stage MST, a four-stage MST with a 1-3-4-5 stage structure, a testlet-based CAT and an item-level CAT. Two P&P tests were also included as baseline conditions; one of the P&P tests was the same length as the MST and CAT conditions, while the other one was twice as long. The modules in the MST conditions contained one to three five-item testlets depending on the stage structure; while the testlet-based CAT adaptively selected five testlets. The CAT conditions also implemented the randomesque procedure (Kingsbury & Zara, 1989) for exposure control. Parameters for items in the MST and CAT conditions were randomly generated within each testlet so that testlets with specific average difficulties were produced. The total test length of 25 items was used in each of the MST and CAT conditions. The P&P conditions used item parameters taken from two intact test sections of Law School Admission Test (LSAT) and had test lengths of 25 and 51 items. All items were measured with the 1PL-IRT model. For all conditions, 25,000 simulated examinees took the test and MAP was used estimated the final examinee

abilities. The study found that, as expected, the item-level CAT was the most accurate and precise in ability estimation, particularly at the extremes of the θ scale. However, the study considered an item-level CAT impractical for many large-scale testing programs such as the LSAT because of its administrative limitations. All testlet-based designs (MST and CAT) yielded acceptable psychometric properties, improved precision over the P&P test of the same length, and similar precision to that of the P&P test of double length.

Patsula (1999) investigated how the various design factors in a MST impacted its performance relative to an item-level CAT and a P&P test. The MST factors manipulated included number of stages (2 conditions), number of modules per stage (2 conditions) and the proportion of total test items in each of the stages (3 conditions), resulting in a total of 12 MST conditions. These were compared against an item-level CAT with the conditional Sympson-Hetter exposure control procedure (Stocking & Lewis, 1998) and a P&P test built to a target information function from a typical CAT with the same maximum conditional exposure rate as the item-level CAT condition. All conditions had overall test lengths of 36 items that were drawn from the same pool of 418 items. The item parameters had been pre-calibrated using the 3PL-IRT model and were from the Logical Reasoning section of the LSAT. A total of 5,000 simulated examinees were administered each condition and MLE was used for estimating ability. The results showed again that the item-level CAT was the most accurate and efficient in its ability estimation while the P&P test was the least accurate and efficient. For the MST test designs, increasing the number of stages increased estimation accuracy; while increasing the number of modules in the later stages increased both estimation accuracy and efficiency. The proportion of total test items in each stage had little impact on estimation at most ability levels. The study also examined the exposure rate of the various test

conditions and found that the item-level CAT performed best with the highest utilization rates and smallest average exposure rates.

Jodoin (2003) was interested in how item pool characteristics, overall test length, and levels of exposure control impacted the psychometric characteristics of LFTs, MSTs, and item-level CATs. Thus, the conditions manipulated by the study included item pool quality (three levels), degree of match between test and item pool content specifications (two levels), test length (two levels), and item exposure levels (several conditions depending on test design). The exposure levels were controlled for the LFTs and MSTs, by the total number of non-overlapping forms or panels assembled; and for the CAT condition, by using either the unconditional or conditional Sympson-Hetter exposure control procedures (Sympson & Hetter, 1985; Stocking & Lewis, 1998). For all three test designs, items were drawn from six simulated item pool of 450 items. The six item pools varied based on the levels of item pool quality and match between test and item pool content specifications. The two levels of test lengths were 40 or 60 items. A random sample of 9,000 examinees was simulated to take each of the test conditions and EAP was used for ability estimation in all test designs. The study found that test reliability, measurement precision and classification precision all increased with higher item pool quality, longer tests, more adequate match between item and test pool content specifications, and less restrictive item exposure levels. The CAT design had the most superior psychometric properties followed by MST and then LFT.

Davis and Dodd (2003) were interested in comparing measurement properties, exposure rates and pool utilization for testlet-based CATs and MSTs. The test designs compared included one MST with the 1-3-3 stage structure and three testlet-based CATs: one with MI selection and no exposure control, one with MI selection and a modification of the within .10 logits (Lunz & Stahl, 1998) exposure control procedure, and one with

random item selection. The tests in each of the four conditions were built from an item pool of 149 passages (testlets) from the Verbal Reasoning section of the Medical College Admission Test (MCAT). Each testlet contained 6, 7, 8 or 10 dichotomous items and the testlets were calibrated using the partial credit model (Masters, 1982). All examinees were administered seven testlets for a total of 55 items. For the MST condition, eight non-overlapping panels were assembled to control for testlet exposure. A total of 1,000 simulated examinees took each of the four test conditions and their abilities were estimated using MLE. The study found that, as expected, the testlet-based CAT with random selection produced the best exposure and pool utilization rates, but the worst measurement properties; while the CAT with MI selection but no exposure control yielded the best estimation accuracy and precision, but the worst exposure statistics. Both the MST and the CAT with the modified within .10 logits procedure performed well in terms of measurement precision and exposure control with the MST condition having the most superior performance when all test-related factors are considered.

Hambleton and Xing (2006) sought to compare the performances of computer-based LFTs, MSTs and CATs used for making pass-fail decisions. The study included a MST condition with five panels using the 1-3-3 stage structure, an item-level CAT condition with the conditional Sympson-Hetter exposure control procedure, and an LFT condition with five non-overlapping forms. The MST and LFT designs were also crossed with two optimality conditions. The item pool consisted of 600 dichotomous items from an existing credentialing exam whose item parameters were calibrated using the 3PL-IRT model. A total of 5,000 examinee abilities were simulated from the standard normal distribution and they were estimated using EAP estimation. The study found once again that the CAT design performed the best in terms of decision accuracy and consistency, followed by MST and then LFT.

STATEMENT OF PROBLEM

From the review of recent literature comparing MST with CAT, it was clear that because of the level of adaptation, item-level CATs consistently provided more accurate and precise ability estimates than MSTs, which in turn, had better psychometric properties than traditional P&P tests or LFTs. The case for MSTs was typically made on the basis of non-psychometric advantages such as more administrative control over content quality and allowing examinees the opportunity to review items (Mead, 2006; Hendrickson, 2007).

However, because research in MST had only begun in earnest recently, numerous properties related to the design of MSTs and their relative merits compared to CAT were still unclear. Exposure control was one such area. Of the literature reviewed, only three studies (Patsula, 1999; Jodoin, 2003; Davis & Dodd, 2003) investigated exposure control as a manipulated condition or item exposure and pool utilization rates as dependent measures. Both Patsula (1999) and Jodoin (2003) found that item-level CATs with *conditional* exposure control procedures performed better than the various MST designs in both measurement characteristics and exposure control. Davis and Dodd (2003) found that a testlet-based MST had slightly superior overall performance compared to a testlet-based CAT with a *randomization* exposure control procedure. Given the concerns about test security due to the frequent exposure of test items, more research in this area was certainly warranted.

Specifically, was the difference in performance found in these studies due to the type of exposure control procedure implemented for CAT? Boyd (2003), in comparing several exposure control procedures for testlet-based CAT, found that Revuelta and Ponsoda's (1998) progressive-restrictive procedure outperformed both randomization and conditional exposure control procedures in terms of exposure rates while producing

comparable measurement properties. How would a testlet-based CAT with the progressive-restrictive procedure, a hybrid of randomization and conditional procedures, perform relative to a testlet-based MST design?

Also, in a recent commentary about multistage testing, Stark and Chernyshenko (2006) provided several suggestions for future research. One conjecture they made, after reviewing several recent MST studies (Luecht et al., 2006; Jodoin et al., 2006; Hambleton & Xing, 2006; Chuah at al., 2006), was that the greatest impact on the ability estimation and classification accuracy of MSTs would lie in the *test length* instead of the other test design considerations. They recommended that future studies include conditions with much shorter test lengths (15-20 items or fewer) to explore the efficiency gained in using an MST over a LFT. For a given item pool, varying the test length of a MST (as well as a CAT) would affect the item exposure and pool utilization rates. As such, it would be worthwhile to investigate not only how test length affected the measurement properties of a MST, but also its exposure control properties relative to a CAT of the same length.

Closely related to test length was the size of the item pool from which a MST or CAT could draw testlets and items. It was reasonable to assume that when the item pool for both types of test designs was smaller, both measurement precision and exposure control properties would be adversely affected because of the reduction in available items. However, would the effect on CATs and MSTs be similar or would one test design be more robust? No MST study had investigated the effect of pool size on its measurement properties. Nor had any study compared the differential effects of pool size on CAT and MSTs.

Additionally, MSTs panels are typically built with specific assumptions about the ability distribution of the underlying examinee population. Modules are hence built to

particular target test information functions and routing methods route examinees according to these assumptions (Luecht & Nungester, 1998). Because a CAT adapts to its examinees, as long as the item pool permits, a CAT should be relatively robust to shifts or changes in the underlying ability distribution. Stark and Chernyshenko (2006), however, wondered how the measurement characteristics of a MST would be affected if there were differences between the actual and assumed ability distribution. All MST studies investigated had simulated examinee abilities (θ) from a standard normal distribution and construct their MSTs based on this normality assumption. How would the psychometric properties and exposure rates of these same MSTs change if the underlying ability distribution were in fact different? Such a scenario could occur in practice if tests were constructed based an inaccurate knowledge about the population; or, if the ability distribution were to change over time due to, for instance, curriculum revision, instructional improvement or educational reform.

Lastly, virtually all research related to testlet-based CATs had implemented tests that adapt between testlets; that is, their level of selection was at the *testlet* level instead of the *item* level. A few studies (Wainer et al., 1991; Wainer et al., 1992) had investigated testlet-based tests that adapt after each item and found little gains in measurement efficiency. However, these studies examined tests with only a single testlet, very short test lengths, and were conducted on hierarchical testlets instead of CATs. Since these studies took place, TRT was proposed as a viable measurement model for testlets. One of the purported advantages of TRT was that it permits *ad hoc testlet construction* in CATs (Wainer, Bradlow & Du, 2000; Wainer et al., 2007). However, no study in the literature investigated had evaluated the improvements in measurement efficiency and exposure rates with such a TRT CAT design. Also, while several MST studies (e.g. Zenisky, 2004; Hendrickson, 2007) suggested using TRT as a measurement

model for MSTs, no studies had implemented a MST using TRT. Thus, it would be interesting to examine and compare TRT-measured MST and CAT designs.

In summary, the goal of this dissertation was to compare three testlet-based adaptive test designs (a CAT that adapts between testlets, a CAT that adapts between and within testlets, and a testlet-based MST) under several manipulated test conditions (test length, item pool size and underlying ability distribution). The three adaptive test designs were implemented using the 3PL-TRT model. Exposure control was enforced in for the two CAT designs by implementing the progressive-restrictive procedure and in the MST by including multiple test forms (or panels). The three designs were compared on measurement accuracy and precision as well as their exposure control properties. Specifically, the research questions addressed by the study included:

1. In general, how do the three adaptive test designs compare in their measurement effectiveness and exposure control properties?
2. Does the total test length have a differential effect on the measurement effectiveness and exposure control properties of the three adaptive test designs?
3. Are the adaptive test designs affected differently by a reduction in the test-length-to-pool-size ratio in terms of their measurement and exposure control properties?
4. What effect does a mismatch between actual and assumed underlying ability distribution have on the measurement and exposure control properties of the MST? How does this effect compare to those of the two CAT designs?

CHAPTER THREE: METHODOLOGY

DESIGN OVERVIEW

In this dissertation, two CAT designs and one MST design were compared across several manipulated test conditions. The two CAT designs compared include a testlet-based CAT that adapts between testlets (testlet-level CAT), and a testlet-based CAT that adapts between items within each testlet (item-level CAT). The MST design had the commonly-used 1-3-3 stage structure, with panels and modules constructed by hand. The test conditions manipulated included two total test length conditions, two item pool size conditions, and two different types of underlying examinee ability distributions. These manipulated conditions were fully crossed with one another and with the three adaptive test designs yielding a total of (3×2×2×2=) 24 study conditions.

Item and testlet parameters for the item pool were estimated using the three-parameter logistic testlet response theory (3PL-TRT) measurement model (Wainer, Bradlow & Du, 2000). The parameter estimates were obtained by calibrating real response data from a large statewide reading examination. Exposure control was implemented within each test design. For the CAT designs, the progressive-restrictive exposure control procedure was used. For the MST design, item and testlet exposure were controlled by constructing multiple panels and by setting limits on the number of times modules and testlets may overlap across panels and modules, respectively.

Maximum information (MI) was used to adaptively select items, testlets or modules to administer in each of the three test designs. The expected a posteriori (EAP) estimation procedure (Bock & Mislevy, 1982) was used for the provisional and final estimations of the ability (θ) and examinee-specific testlet effect (γ) parameters in all three test designs. A fixed-length stopping rule was used for all test designs, with the

number of items on the test determined by the total test length condition. Indices of measurement accuracy and precision, exposure rates, pool utilization as well as item and testlet overlap were used to compare the three test designs across the study conditions.

ITEM POOL

The testlet and item parameters used in this dissertation were estimated by calibrating real student response data from an existing large-scale assessment. The data were obtained from recent administrations of a statewide reading examination at one particular grade level in a southwestern US state. This reading examination is given annually to all students in the state as part of state and federal accountability requirements. The statewide dataset obtained contained exam data from three consecutive school years. In the first year, 46 operational test forms were given to a total of 137,433 students at this particular grade level, resulting in an average of 2,988 students per form with a minimum of 2,656 and a maximum of 3,040 students per form. In the second year, 43 test forms were administered to a total of 125,314 students at this grade level, with an average of 2,914, minimum of 2,666, and maximum of 2,956 examinees per form. In the third year, a total of 39 operational test forms were given to 122,825 students at the grade level. An average of 3,149 students took each test form, with a minimum of 3,083 and a maximum 3,904 examinees per form. In all three school years, every test form consisted of a total of 52 dichotomously-scored multiple-choice items from 5 reading passages (or testlets). Each reading passage on the operational test forms had either 10 or 12 associated items.

Of the 52 items in each test form, 42 of them (representing 4 testlets) were common across all test forms and are known as the *base-test items*. The base-test items had been previously field-tested and reviewed by curriculum experts and educator committees consisting of representative teachers and content specialists from across the

state. As a group, the base-test items satisfied the content specifications and statistical targets described in the test blueprint for this reading exam. Each student's performance on the 42 base-test items counted towards their final exam score.

The remaining 10 items (representing 1 testlet) were unique to each test form and are called the *field-test items*. While the content of these items had also been reviewed by educators and specialists, they had not previously appeared on any exams. The field-test items were included in the exam for the sole purpose of gathering information about the psychometric properties of the items under an operational setting. Thus, each student's performance on the 10 field-test items did not count towards their final exam grade. It should be noted, however, that the 10 field-test items were embedded into the operational test forms. Examinees were unable to distinguish the base-test items from the field-test items. Thus, no differential motivation effects were expected to be in the students' responses to the base-test and field-test items.

The item pool for this dissertation contained both base-test and field-test item and testlet parameters. In the first school year, a total of 18 unique passages (4 base-test + 14 field-test passages) with 397 unique base-test and field-test items were tested across the 46 operational test forms; in the second year, 18 unique passages (4 base-test + 14 field-test passages) with 364 unique base-test and field-test items were tested across the 43 forms; and in the third year, 18 unique passages (4 base-test + 14 field-test passages) with 334 unique items were distributed across the 39 forms. The four passages and associated items that compose each year's base test were typically selected from the set of items that were field-tested in the previous school year. For example, four of the passages that were field-tested in the first year were used in the base test for the second year. As such, these 4 passages along with the 42 items associated with them appear in the datasets for both the first and second school years. Similarly, another four passages and their 42 associated

items overlapped in the datasets obtained for the second and third school years. Thus, summing across the three school years, a grand total of up to 1,011 unique reading items representing 46 passages were available from the obtained statewide datasets to include in the item pool for this dissertation.

One important distinction should be noted about the base-test and field-test passages. Passages in the base test had considerably fewer items associated with them than the passages in the field test. This is because an identical set of base-test items and passages were included across all test forms in each school year so that examinees were evaluated on the same set of items regardless of the test form they were administered. Each test form, however, had a different set of field-test items and reading passages. While two or more test forms could share the same field-test reading passage, different sets of items associated with a given passage typically appeared on each of the test forms. This was done for this statewide assessment so that if a particular field-tested passage were selected to be part of a future base test, the test constructor would have a larger group of associated items to choose from for that passage. Consequently, each of the 42 field-test passages from across the three school years had between 18 to 32 associated items; while the 12 passages on the base tests have only either 10 or 12 associated items. Eight of these base-test passages (for the second and third school years), however, did have additional items that were field-tested across the test forms in the earlier year. Thus, only the four base-test passages from the first year had smaller sets of (10 or 12) items to choose from in the item pool.

PARAMETER ESTIMATION

The items parameters were originally estimated using the 1PL-IRT (or Rasch) measurement model (Rasch, 1960; Wright & Stone, 1970) for the statewide assessment. This implied that any dependencies between items associated with the same passage were

not accounted for in the original calibration of the item responses. In addition, all items were assumed to be equally discriminating with no chance of answering correctly through guessing. In this dissertation, these assumptions about discrimination and guessing were relaxed and the local dependency within testlets was accounted for by re-estimating the item parameters using the three-parameter logistic testlet response theory (3PL-TRT) measurement model.

Under the 3PL-TRT model, the item and testlet parameters that needed to be estimated include for each item *j*, the discrimination (a_j), difficulty (b_j), pseudo-guessing (c_j) parameters; and for each testlet, *d(j)*, the variance of the testlet effect ($\sigma^2_{d(j)}$). The item responses in the real dataset were calibrated with the SCORIGHT software program (Wang, Bradlow & Wainer, 2001). As it is a unique feature of the 3PL-TRT model, the testlet effect was allowed to vary across testlets so that reading passages can exhibit different degrees of local dependency among its associated items. SCORIGHT estimated parameters using a Markov Chain Monte Carlo (MCMC) technique with Gibbs sampling (Geman & Geman, 1984). Characteristics of the posterior distributions for the item and testlet parameters obtained using the Gibbs sampler were then used to estimate the corresponding parameter values. The statewide assessment data was calibrated using a total of 8,000 MCMC iterations. The first 7,000 iterations served as a burn-in period, during which the posterior distributions could stabilize. The values from the burn-in iterations were dropped in the final estimation of the posterior distributions. Every fifth-iteration of the final 1,000 iterations was used to create the posterior distributions of the item and testlet parameters. This procedure was similar to what Boyd (2003) did to estimate the 3PL-TRT item and testlet parameters in her study.

Items and testlets parameters from the 46, 43 and 39 different test forms in the three school years were estimated separately using SCORIGHT. This yielded a total of

(46+43+39=) 128 separate 3PL-TRT calibration runs. The resulting item and testlet parameters were combined to form the initial item pool. If any item or testlet appeared in multiple test forms within or across school years, then the parameter estimates obtained using the largest number of student responses were used.

Upon examining the parameter estimates from the initial calibration runs, six problematic items were identified. These items had parameter estimates that were deemed unstable – they had either a discrimination (a) parameter that was greater than 3.0 or a difficulty (b) parameter that was less than -4.0, and the standard error for the parameter in question was greater than 0.30. Thus, item responses for these six items were removed from any of the test forms that they were a part of. The affected test forms were then re-calibrated in SCORIGHT to remove any effects the problematic items may have had on their original calibrations. This resulted in a re-calibration of 45 of the 128 test form. The high number of re-calibrated test forms was due to the fact that one of the problematic items was a base item in the second school year. As such, all 43 test forms in that year needed to be re-calibrated.

After the re-calibration, all item parameter estimates were examined again and found to be stable. Thus, the final full item pool consisted of a total of 1,008 items associated with 46 testlets. Figure 3.1 gives the distribution of test information across the proficiency (θ) scale in the final item pool. Appendix A also includes figures with distributions of the a, b and c parameters in pool (Figures A1, A2 and A3 respectively).

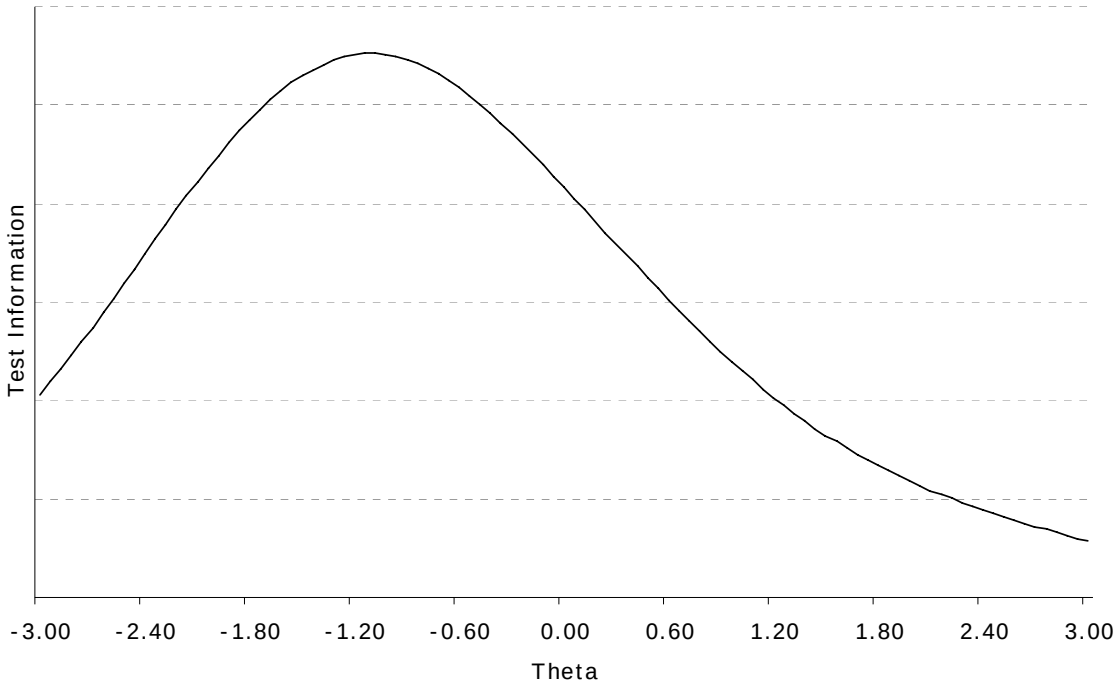

Figure 3.1: Distribution of test information in the final item bank

Figure 3.1 shows that the test information in the final item pool is positively-skewed. More pool items had lower item difficulties (see Figure A2) resulting in a test information function that peaked at around -1.0. This skewed test information distribution, while unintended, had implication on the measurement precision of the various adaptive test designs (as shown later).

Figure 3.2 below shows the distribution of the proficiency (θ) estimates that resulted from the 3PL-TRT calibrations across the three school years. A total of 382,916 θ estimates are represented in this figure. And it shows that the θ estimates were distributed symmetrically and approximately normal. Descriptive statistics of the θ estimates also show that the mean of the distribution was approximately zero (mean of θ = -0.001) with a standard deviation close to one (standard deviation of θ = 0.924).

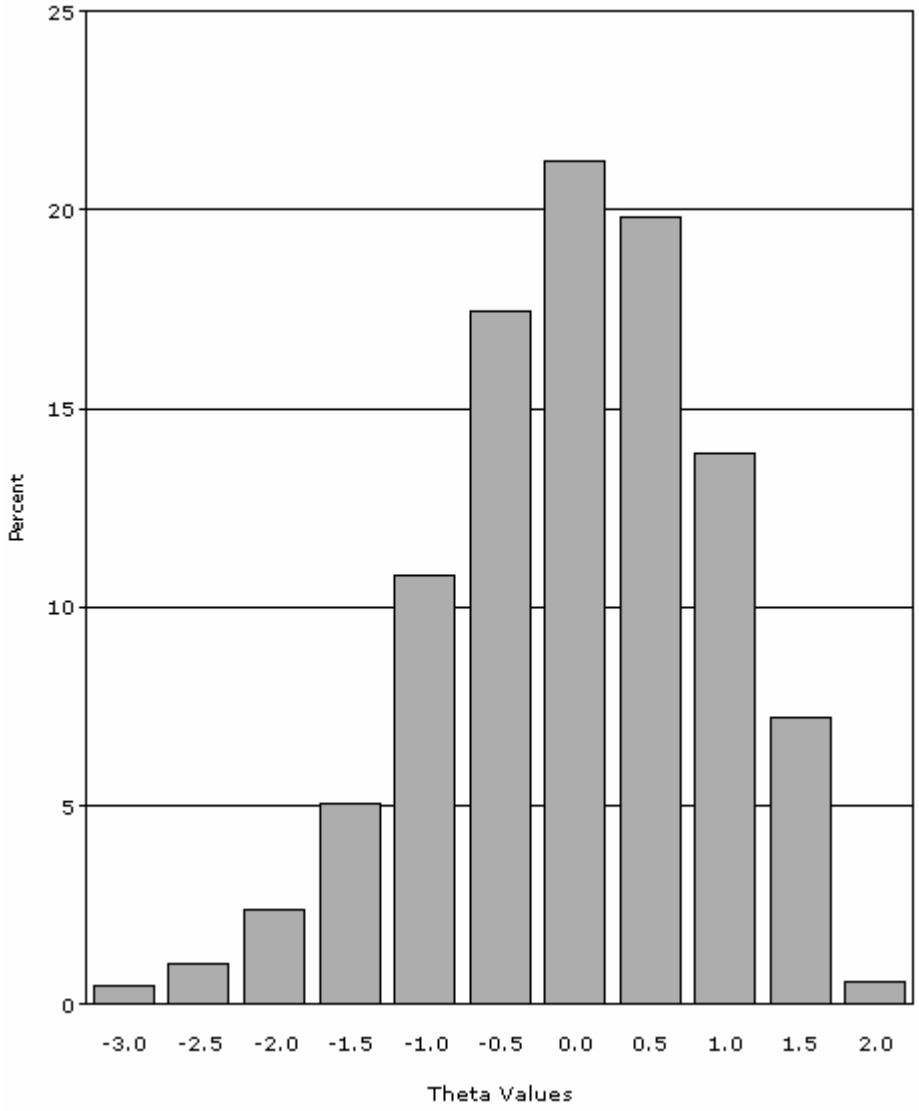

Figure 3.2: Distribution of estimated thetas across the three school years

The reason that the test information for the item pool was positively-skewed while the estimated theta values were approximately normally distributed was an artifact of the TRT calibration software. Unlike several Rasch (1PL IRT) calibration programs which center their scales on items, SCORIGHT centers its scale on people. Consequently, examinees with the mean proficiency estimates set the origin of the scale, and the

estimated item difficulty parameters were defined relatively to that origin. The distribution of the estimated theta values, however, did validate the decision in this dissertation to generate examinee proficiencies from the standard normal distribution as one of the underlying ability distribution conditions.

MANIPULATED CONDITIONS

Test Length

Two total test length conditions were simulated in this dissertation: long and short. Tests under the *long* test length condition consisted of 42 items. This was equal to the length of the base-test for the statewide reading assessment on which the item and testlet parameters were based. Also, like the statewide reading assessment, the long test length condition included 4 reading passages or testlets. One testlet had 12 associated items while the other 3 testlets had 10 items each.

Tests under the *short* test length condition contained 21 items. This was half the test length of the statewide reading assessment exam. One decision that needed to be made about the short test length condition was whether to reduce the total number of testlets within each test (while maintaining the same number of items per testlet) or to reduce the number of items per testlet (while maintaining the same number of testlets within each test). For this dissertation, the total number of passages within each test stayed constant across long and short tests while the number of items within each passage was halved for the short test length condition. This meant that each short test still consisted of 4 testlets, but one testlet included only 6 items while the other 3 testlets had only 5 items each.

One of the advantages of both CATs and MSTs was that they are able to achieve the same degree of measurement precision with shorter test lengths. Also, Stark and

Chernyshenko (2006) in reviewing several MST studies have conjectured that the greatest impact on the measurement precision of MSTs lied in the test length instead of other test design considerations. They recommended that future studies include conditions with much shorter test lengths (15-20 items) to test this hypothesis. Only one MST study in the literature review included test length as a manipulated condition (Jodoin, 2003). Thus, including test length as a manipulated condition allowed this study to assess the impact on measurement precision when the test length of a CAT or MST was halved. While one would expect the measurement precision to be worse for the shorter test length, the loss in precision should be minimal if the stated advantage of CATs and MSTs were true. The study findings would also allow the comparison of differences in precision loss and change in item pool utilization rates across the three adaptive test designs.

Pool Size

Two pool size conditions were simulated in this dissertation: full and reduced. The *full* pool size contained the entire set of testlet and items available from the statewide reading assessment. Thus, it included a total of 1,008 items associated with 46 testlets.

The *reduced* pool consisted of about two-thirds of the total testlets in the full item pool and therefore approximately two-thirds of the number of items available from the statewide reading assessment. To scale down the item pool, a process similar to what was performed in a CAT study by Koch and Dodd (1989) was utilized. First, the distribution of test information across the proficiency (θ) scale for the full item pool was examined (see Figure 3.1). The goal was to create a reduced item pool with a test information function that had a similar shape, but inevitably less total information across the θ scale. This was done by computing the *testlet* information (that is, the sum of item information provided by all items in a testlet) of the 46 testlets in the full item pool. Then, the testlets

were sorted by where the peaks of their testlet information function were and divided into 15 groups, each consisting of three testlets. Within each group of threes, the testlet with the least amount of total testlet information was then removed from the pool. One testlet – the one whose mode was at the highest θ value – was not assigned to a group of three and was automatically included in the pool. Thus, this procedure yielded a reduced item pool of 31 testlets, with a total of 741 available items. Figure 3.2 gives the resulting test information function for the reduced item pool and compares it with that of the full item pool. The similarity in the shapes of the two test information functions shows that the procedure used to scale down the full item pool achieved its objective.

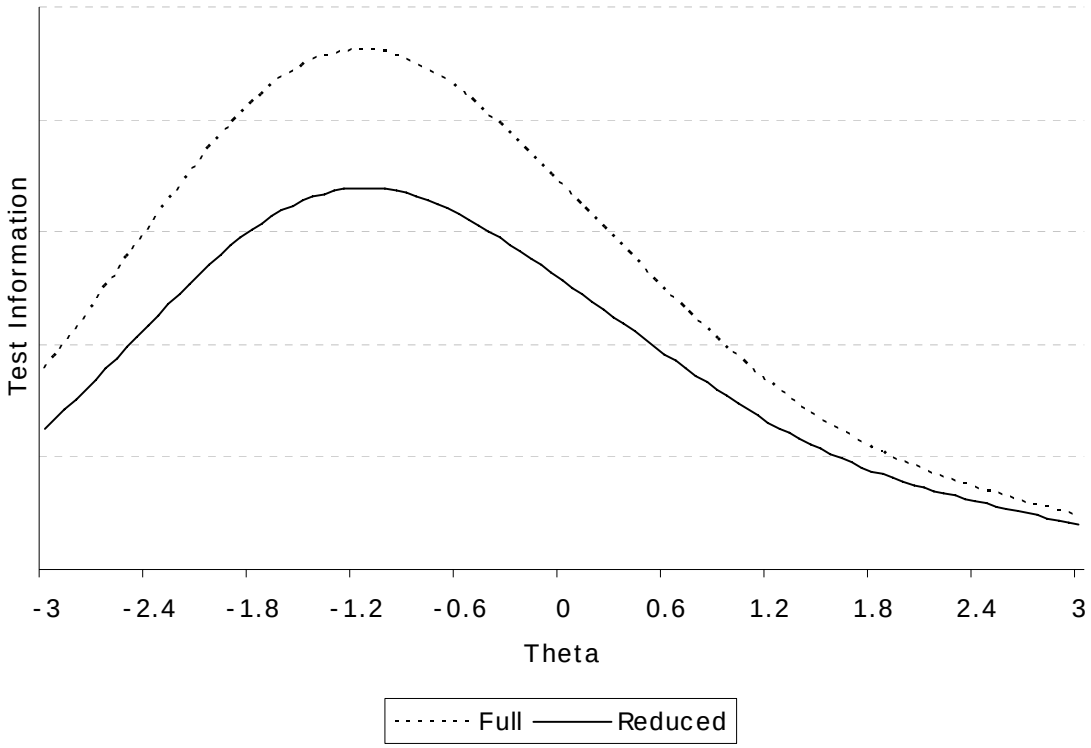

Figure 3.2: Test information functions for the full and reduced item pools

The reason for choosing a reduced pool size that was approximately two-thirds (as opposed to one-half) of the full pool size was to allow for variation in the ratios of *test*

length to pool size. Recall that the two total test length conditions had either 21 items (short) or 42 items (long). If the two pool size conditions also had a 1-to-2 ratio, then the *short* test length condition combined with the *reduced* pool would have had the same length-to-pool-size ratios as the *long* test length condition combined with the *full* pool. The former combined condition would therefore be simply a proportionally scaled-down version of the latter one, and the exposure properties for these two combined conditions would likely not be as informative. The choices of test length and item pool size conditions for this dissertation, however, led to combined conditions with systematically different test-length-to-pool-size ratios. These ratios are shown in Table 3.1.

Table 3.1: Approximate test-length-to-pool-size ratios for the study conditions

	Test Length Conditions	
Pool Size Conditions	Long	Short
Full	1:24	1:48
Reduced	1:18	1:36

If the pool of items and testlets that a CAT or MST can draw from were reduced, then, holding all other factors constant, the measurement precision of the test would be expected to decrease while the exposure and pool utilization rates would likely increase. The question of interest for this study was whether there would be a differential effect across the three adaptive test designs compared in this study. Would one of the test designs be less impacted by a reduction in pool size than the others? An additional issue that can be investigated, as illustrated in Table 3.1, was the interaction effect of test length and pool size. Would the three test designs perform differently as the test-length-to-pool-size ratio changed?

Ability Distribution

There were two underlying population ability distribution conditions in this dissertation: normal and skewed. Under the *normal* distribution condition, the simulated examinee ability parameters (θ) were sampled from a normal distribution with a mean of zero and standard deviation of 1. This ability distribution represented what is typically assumed for the examinee population in most operational settings and research studies. It also matched the distribution of θ estimates from the SCORIGHT-calibrated statewide reading assessment data.

Under the *skewed* distribution condition, examinee abilities were first sampled from a beta distribution with $\alpha = 5.0$ and $\beta = 1.8$. This resulted in a negatively-skewed distribution with a mean of .74, standard deviation of .16, skew of -.73 and kurtosis of zero. The sampled θ values were then transformed so that the distribution was centered on zero and had a standard deviation of 1. The resulting ability distribution had a mean of approximately 1.5. This procedure of sampling from a negatively-skewed distribution was used in Gorin, Dodd, Fitzpatrick and Shieh (2005). Such a distribution would be characteristic of an test-taking population that was, on average, more proficient in the trait measured by the test. This can occur in practice if, for example, after several years into an assessment program, instructional improvement and familiarity with the exam format led to growth in student achievement on the test. Figure 3.3 shows the distribution of the two underlying ability conditions simulated in this study.

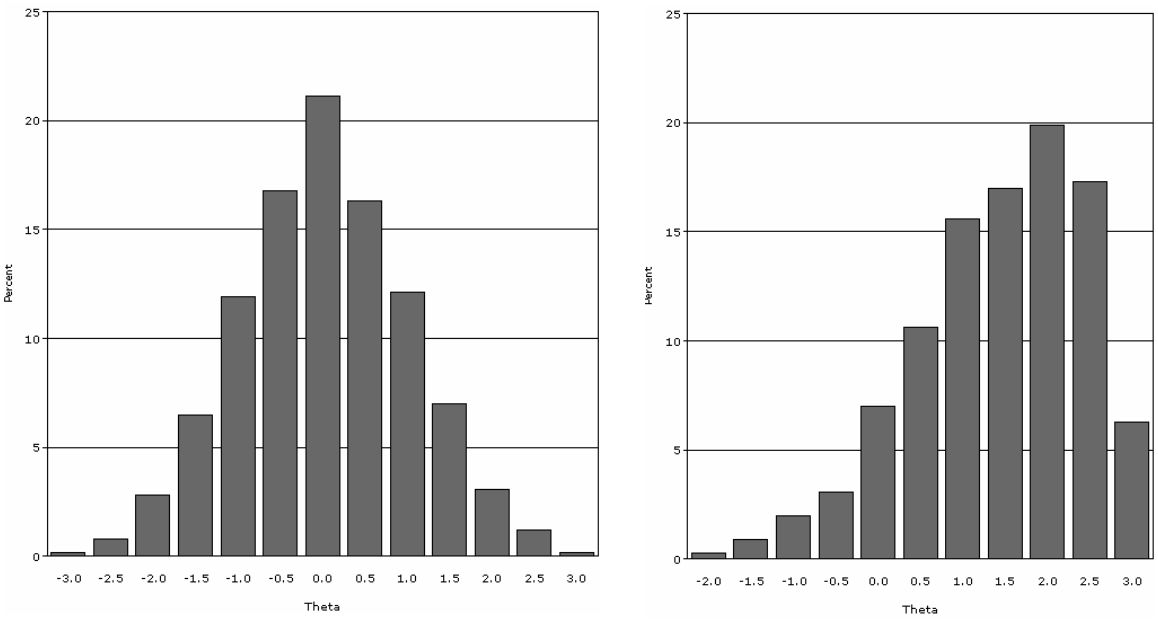

Figure 3.3: Distribution of θ values for the normal and skewed conditions

The reason for including this condition was because all MST studies and most CAT studies reviewed in the literature investigation assumed normality for the underlying population ability distribution. The impact of the underlying ability distribution on the estimation effectiveness of a CAT was found to be minimal (Chen, Hou, Fitzpatrick & Dodd, 1997; Gorin et al., 2005). Its effect on the exposure control properties of a CAT was also expected to be small relative to a MST, as long as the item pool was sufficiently large. This is because a CAT algorithm would generally adapt the items it administers to whatever provisional ability estimate it has for the current examinee. The underlying ability distribution should therefore have little bearing on the operation of a CAT. In contrast, a MST is assembled prior to its administration and test constructors need to make specific assumptions about the underlying ability distribution. Under these assumptions, modules would be built to satisfy particular test information targets (as was done in this study). Routing methods also direct examinees between

stages in ways that would lead to relatively even exposure of modules within each stage. Stark and Chernyshenko (2006), however, wondered how the characteristics of a MST would be affected if there were a mismatch between the assumed and actual underlying ability distributions. What impact would this have on the measurement properties of the MST? How would this affect the exposure rates across its modules? And how would the MST's measurement and exposure properties compare to those of the CAT designs?

DATA GENERATION

The data generation process involved two steps. First, samples of examinee-specific parameter values were generated. Then, response data for every item and for each examinee were simulated based on the examinee-specific parameters and the SCORIGHT-calibrated item and testlet parameters. The simulated response data were then used across the study conditions in this dissertation. Details of how the examinee-specific parameters and response data were generated are given next.

There were a total of 10 replications for this study. Within each study replication, two samples of 1,000 simulated examinee ability (θ) values were randomly drawn from a probability distribution specified by the underlying ability distribution condition. Recall that the two ability distribution conditions were normal and skewed (see Figure 3.3). Thus, one sample, called the *normal sample*, had 1,000 θ values generated from the standard normal distribution; the other sample, called the *skewed sample*, contained 1,000 θ values drawn from a negatively-skewed beta distribution. Thus, a total of 2,000 θ values were generated in each replication, and these were the known ability (θ) values for the examinees. Also, under the 3PL-TRT measurement model, each examinee (*i*) should have a person-specific testlet effect parameter ($\gamma_{id(j)}$) associated with each testlet, *d(j)*. Within each testlet, the γ parameter is assumed to normally distributed with a mean of zero and a variance equal to the variance of the testlet effect ($\sigma^2_{d(j)}$) for the testlet.

(Wainer et al., 2000). The testlet effect variance ($\sigma^2_{d(j)}$) was one of the parameters estimated by SCORIGHT for each testlet in the item pool during its calibration of the real dataset. So, an examinee's γ parameter for each testlet was simply randomly generated from a $N(0, \sigma^2_{d(j)})$ distribution. This was done separately for the two samples (normal and skewed) in each replication. Because there were 14 testlets in the full item pool, this implied that 14,000 γ values were generated for the normal sample and another 14,000 γ values were generated for the skewed sample, resulting in a total of 28,000 simulated γ parameters per replication. These are the known γ values for each examinee and each testlet. Note that for the reduced item pool conditions, only the γ values associated with the 31 selected testlets were used.

Next, a modification of the 3PL-TRT SAS data generation program used by Boyd (2003) was used to generate response data. For each simulated examinee (i), the probability of responding correctly to a particular item (j) was computed using Equation (5) and is based on the examinee's θ_i value, the item's discrimination (a_j), difficulty (b_j) and pseudo-guessing (c_j) parameters and the person-specific testlet effect parameter ($\gamma_{id(j)}$). To introduce random error, this probability was compared against a randomly generated value between 0 and 1 from the uniform distribution. If the random value was less than or equal to the probability, then the examinee received a correct response (1) for the item; if the random value was greater than the probability, then the examinee was assigned an incorrect response (0). This was done for every item and for each person in the simulated examinee samples (normal and skewed) within each replication.

The data generation steps described above was repeated for 10 replications, resulting in a grand total of 20 samples (10 normal samples and 10 skewed samples) of examinee parameters and item responses for this dissertation. Recall also that a total of 24 study conditions will be examined and, within a replication, the same examinee

sample was given to all the study conditions with the corresponding underlying ability distribution. So, within each replication, the normal sample was used in the 12 study conditions with the normal ability distribution; while the skewed sample was used in the 12 study conditions with the skewed ability distribution.

CAT SIMULATIONS

The CAT simulations were based on modifications to a SAS program that was originally created by Chen, Hou and Dodd (1998), modified by Davis and Dodd (2003), and then further modified by Boyd (2003). The program was modified to simulate the two testlet-based CAT designs: the testlet-level CAT and the item-level CAT.

Common CAT Design Components

For both CAT designs, expected a posteriori (EAP) estimation was used for provisional and final estimations of the examinee ability (θ) and person-specific testlet effect parameter (γ). The EAP estimation procedure implemented was based on a normal prior with 30 evenly-spaced quadrature points along the ability scale ranging from -4 to +4. The quadrature points were used to compute the weights for determining the posterior distribution of the estimated parameter (θ or γ). Fixed-length stopping rules were used with test lengths of either 21 or 42 items, depending on the total test length condition. For exposure control, both CAT designs implemented the progressive-restrictive procedure (Revuelta & Ponsoda, 1998). The maximum exposure rate at the testlet level was set to .30, as Boyd (2003) found this to be a reasonable maximum exposure rate for testlet-based CATs measured by the 3PL-TRT model.

It should be noted, however, that for the testlet-level CAT, exposure control only needed to be implemented at the testlet level. Under this CAT design, the items administered with a testlet were pre-determined. Thus, the exposure control procedure at

the testlet level effectively controls the exposure rates of the items associated with each testlet, too. In contrast, for the item-level CAT design, the items administered with each testlet were determined on the fly during the test. Thus, while the testlet-level exposure control procedure does set an upper bound for proportion of time an item can be administered, an additional exposure control procedure could be implemented at the item level to further control the item pool utilization. This is done for the item-level CAT in this dissertation – the progressive restrictive procedure is implemented at the item level with a maximum exposure rate of .25. This maximum exposure rate was chosen because the maximum exposure rate was already .30 at the testlet level. Thus, a lower maximum was needed at the item level for the progressive restrictive exposure control to have an effect at the item level. An item-level maximum exposure rate of .20 was initially implemented, but it led to convergence issues in the reduced item pool and skewed ability distribution conditions. Thus, the item-level maximum exposure rate was raised to .25.

The distinction in exposure control implementations between the testlet-level and item-level CAT designs was just one of many components that were different between the two CAT designs. Further details about the unique CAT components in each CAT design are given in the following two sections.

Testlet-Level CAT Design

For the testlet-level CAT, adaptation occurred between testlets. This means that once a testlet was selected for administration, a specific set of items associated with the testlet were given regardless of how the examinee was performing on the items within the testlet. In the item pool for this dissertation, almost all testlets had more items associated with them than need to be administered. Thus, to simulate the testlet-level CAT, the set of items that were always administered together with each testlet needed to be pre-selected. For the study conditions with the long (42-item) test length, this meant creating

testlets with 10 pre-selected items and testlets with 12 pre-selected items. For the short (21-item) test length, testlets were created with either 5 or 6 pre-selected items.

However, if only either 5 or 6 items (for the short tests), or 10 or 12 items (for the long test) associated with each of the testlets were chosen to always be administered when the testlet was selected, then a large portion of the item bank would have no chance of every being administered. Such a setup would effectively skew the overall utilization rate of the item pool, yielding an unreasonable and artificially high percentage of items never administered. Thus, to give all items in the pool a chance to be administered, multiple versions or *permutations* of items associated with the same testlet were created.

The permutations were created with two goals in mind. First, the various permutations of items for a testlet needed to be similar in statistical characteristics. In other words, the permutations were parallel forms and were almost interchangeable within the testlet. Second, every item associated with the testlet should be assigned to at least one of the permutations. The number of permutations an item appeared in should also be roughly the same across items to facilitate better pool utilization. To accomplish these goals, a heuristic was devised to determine the number of permutations for each of the 46 testlets and the way items were assigned to each permutation. The number of permutations, P, needed for each testlet was found with the formula,

$$P = \text{ceil}\left(\frac{\text{Number of items available for the testlet}}{\text{Number of items administered with each testlet}}\right) \qquad (16)$$

where ceil(•) is the ceiling function that rounds its inputted argument up to the nearest integer value. The k^{th} item in the p^{th} permutation is then determined as,

$$i = \text{round}\left(p + k \times \left(\frac{\text{Number of items available for the testlet}}{\text{Number of items administered with each testlet}}\right)\right) \qquad (17)$$

where i is the index of the testlet item after all items were sorted by their difficulty (b) parameters (i.e. the i^{th} most difficult item among the available items for the testlet) and round(•) rounds its argument to the nearest integer using the usual rounding rules. If i in Equation (17) comes out to be larger than the number of items available for a given testlet, then it is set to the largest item index.

To illustrate the heuristic, consider a testlet in the item pool with 32 available items. Based on Equation (16), the number of 12-item permutations (used in the long test length condition) needed for this testlet was P = ceil(32/12) = ceil(2.6667) = 3. The 32 items were sorted in ascending order by their b parameters and assigned an index (i) based on this sort order. Then, applying Equation (17), the three 12-item permutations for this testlet were defined according to Table 3.2. Each entry in Table 3.2 is the item index (i) based on the difficulty sort order.

Table 3.2: Example 12-item permutations for a testlet with 32 available items

	Permutation		
Item	P1	P2	P3
I1	1	2	3
I2	4	5	6
I3	6	7	8
I4	9	10	11
I5	12	13	14
I6	14	15	16
I7	17	18	19
I8	20	21	22
I9	22	23	24
I10	25	26	27
I11	28	29	30
I12	30	31	32

This heuristic was applied to create permutations for each testlet in the pool. This was done separately for every combination of test length × item pool size conditions. Consequently, a total of 219 permutations (of 10 items and 12 items) were created for the long test × full pool size condition, 164 permutations (of 10 items and 12 items) were created for the long test × reduced pool size condition, 407 permutations (of 5 items and 6 items) were created for the short test × full pool size condition, and 297 permutations (of 5 items and 6 items) were created for the short test × reduced pool size condition.

Figure 3.4 below gives the testlet information functions for the three 12-item permutations (in Table 3.2) created for the testlet with 32 available items. The figure shows that the resulting permutations were similar in the information they provided (i.e. parallel forms). Thus, the permutations could be given interchangeably whenever this testlet was chosen by the CAT algorithm for administration.

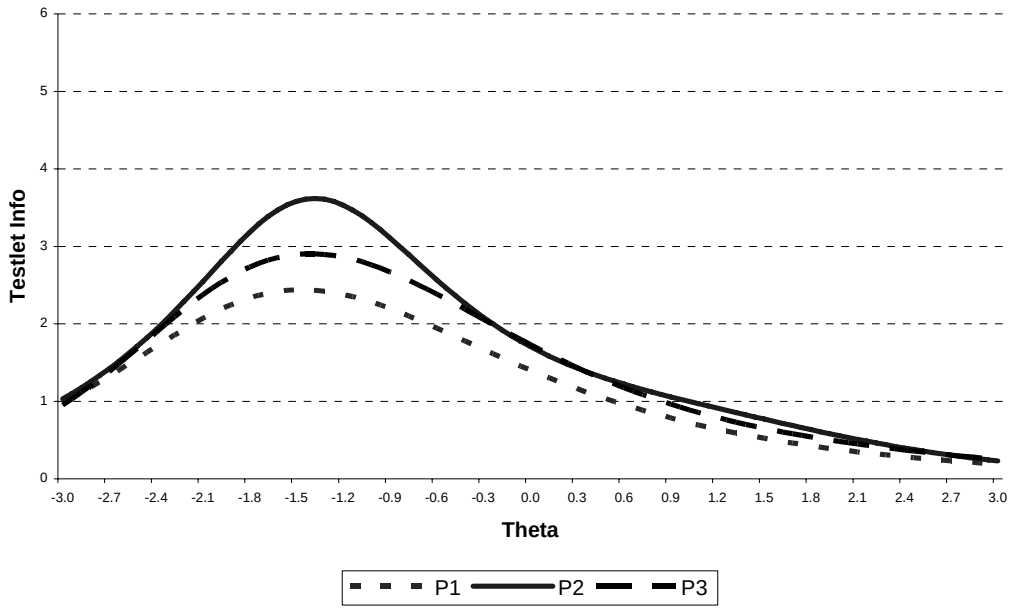

Figure 3.4: Testlet information functions for the 12-item permutations

With these details in mind, the following were the steps in the algorithm for simulating the testlet-level CAT:

1. The initial ability estimate ($\hat{\theta}$) was set to zero.
2. A testlet was chosen for administration from the pool based on $\hat{\theta}$ using maximum information (MI) selection with the progressive-restrictive exposure control procedure (maximum exposure rate = .30) applied at the testlet level only.
3. One of the 10-item (or 5-item) permutations for the testlet was randomly selected for administration. The testlet was marked as administered, so that it could not be chosen for the same examinee again.
4. Provisional ability ($\hat{\theta}$) and testlet effect ($\hat{\gamma}$) parameters were estimated using EAP based on the 3PL-TRT model.
5. Steps 2 to 4 were repeated to select three additional testlets. The only difference was that for the third testlet, one of the 12-item (or 6-item) permutations for the selected testlet was randomly chosen for administration.
6. The final ability and testlet effect parameters were estimated with the entire set of responses using EAP estimation based on the 3PL-TRT model.

Note that in Step 2, MI selection was based on the testlet information, the sum of the item information within the testlet. At that point in the algorithm, the testlet information was computed using only the 3PL-IRT item parameters (a, b, and c) of the associated items because the examinee-specific testlet effect parameter (γ) could not be estimated until items in the testlet had been administered to the examinee. Only after a testlet had been completely administered, could the ability and testlet effect estimates and their standard errors be re-calculated based on the 3PL-TRT model (i.e. in Steps 4 and 6).

Item-Level CAT Design

The testlet-based item-level CAT design functioned very similarly to one that was not testlet-based. That is, it adapts after each item. The major distinction was that the set of candidate items that could be administered was restricted by the current testlet, in addition to any restrictions imposed by the exposure control procedure. Thus, the steps involved in simulating the item-level CAT included:

1. The initial ability estimate ($\hat{\theta}$) was set to zero.
2. A testlet was chosen for administration from the pool based on $\hat{\theta}$ using MI selection with the progressive-restrictive exposure control procedure (maximum exposure rate = .30). Note that MI selection in this step was based on *testlet* information and the exposure rates of the available *testlets* were used by the progressive-restrictive procedure. The testlet was marked as administered, so that it could not be chosen for the same examinee again.
3. The next item to administer was selected from the set of items associated with the chosen testlet. It was also selected based on $\hat{\theta}$ using MI selection with the progressive-restrictive exposure control procedure. However, in contrast to step 2, MI selection in this step was based on *item* information and the exposure rates of the available *items* for this testlets were used by the progressive-restrictive procedure (maximum exposure rate = .25). The item was marked as administered, so that it could not be chosen for the same examinee again.
4. Provisional ability ($\hat{\theta}$) and testlet effect ($\hat{\gamma}$) parameters were estimated using EAP based on the 3PL-TRT model.
5. Repeat steps 3 and 4 until the required number of items were administered for the current testlet. For the third testlet, either 6 or 12 items were required (depending

on the test length condition); while for the other three testlets, either 5 or 10 items were needed.

6. Repeat steps 2 to 5 until four testlets were administered.
7. The final ability and testlet effect parameters were estimated with the entire set of responses using EAP estimation based on the 3PL-TRT model.

Note that in Steps 2 and 3, the testlet and item information used to select the next testlet and item respectively was based on only the 3PL-IRT item parameters (a, b and c). As in the testlet-level CAT, the examinee-specific testlet effect parameter (γ) could not be estimated until the testlet items were all administered to the examinee. Thus, the ability and testlet effect estimates and their standard errors were recalculated based on the 3PL-TRT model only after each testlet were administered completely (i.e. in Steps 4 and 7).

MST SIMULATIONS

Test Structure

The MST condition included 8 panels. The number of panels in a MST directly influences the exposure rates of modules within each panel. The choice of 8 panels for this dissertation meant that on average, the first-stage module of each panel would be administered to 12.5% of the examinees; while modules in the later stages would be administered less frequently. This was therefore equivalent to having a maximum exposure rate of .125, which was less than half of the .30 maximum exposure rate used for the progressive-restrictive procedures in the two CAT designs. It should be noted, however, that .125 is the maximum exposure rate for *modules*. Testlets and items could be shared among modules within or across panels. A testlet that was, for example, shared by two first-stage modules would on average be seen by 25% of the examinees. Thus, to ensure that the maximum exposure rates of testlets and items would be no worse in the

MST condition than they were in the CAT conditions, the rule that no testlet could appear in more than two modules was enforced during the MST module assembly process.

Each panel had the 1-3-3 stage structure: the first stage contained 1 module, and the second and third stages contained 3 modules each. This was a common MST design employed in both literature and practice (e.g. Luecht & Nungester, 1998; Davis & Dodd, 2003; Luecht, Brumfield & Breithaupt, 2006; Jodoin, Zenisky, & Hambleton, 2006; Hambleton & Xing, 2006; Chuah, Drasgow & Luecht, 2006). Like the two CAT designs, the MST administered four testlets to each examinee for a total of either 21 or 42 items (depending on the test length condition).

A decision needed to be made about how the four testlets and the number of items would be distributed across the modules in the three stages. Several arrangements were possible. For this dissertation, the arrangement for the long (42-item) test length conditions was to include two 10-item testlets in the first-stage module, one 12-item testlet in each of the second-stage modules, and one 10-item testlet in each of the third-stage modules. The short (21-item) test length condition had half the number of items within each testlet. That is, there were two 5-item testlets in the first-stage module, one 6-item testlet in each of the second-stage modules, and one 15-item testlet in each of the third-stage modules. Figure 3 shows the test structure of the 42-item MST condition in this dissertation.

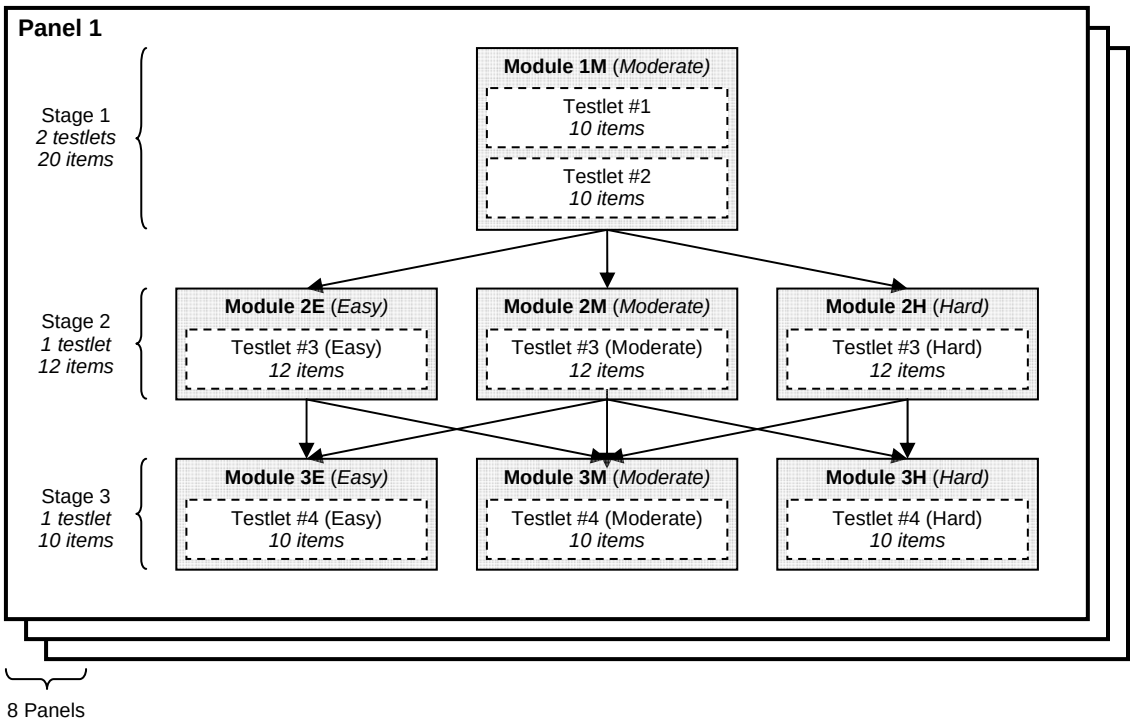

Figure 3.5: The 1-3-3 stage structure for a 42-item MST

The result of this arrangement was that the first-stage module was approximately twice as long as the second- or third-stage modules. The rationale for a longer first-stage module was based on the findings by Loyd (1984) and by Kim and Plake's (1993) that the length of the routing test had a significant impact on the measurement precision of an MST. Administering a longer first-stage module provided more information in the second-stage routing decision. This same rationale was also behind why the second-stage modules were slightly longer than the third-stage modules. This arrangement also made the order in which 10-item (5-item) or 12-item (6-item) testlets were administered analogous to those of the two CAT designs.

Note also from the MST design in Figure 3.5 that between stage 2 and 3, it was not possible for an examinee to be routed from an easy stage-two (2E) module to a hard stage-three (3H) module; nor could an examinee go from a hard stage-two (2H) module

to an easy one in Stage 3 (3E). It was unlikely by chance that an examinee's ability estimate would change from being at one end of the ability (θ) scale to the other after taking the items in a single stage. Thus, disallowing these routes was commonly done in MST design to prevent any aberrant results that would come out of such inconsistent response patterns (Luecht et al., 2006). It could also prevent any negative psychological impact on the examinee that might occur from jumping from easy to hard, or hard to easy, items (Davis & Dodd, 2003).

To route an examinee between stages, a method similar to maximum information (MI) selection in CAT was implemented. This method required computing the test information function for each of the modules in the next stage. For this dissertation, the test information functions for each of the second- and third- stage modules were simply the *testlet* information function of the one testlet in each module. Using the test information functions, the amount of information provided at the provisional ability estimate ($\hat{\theta}$) could be calculated for each module in the next stage. The examinee was then routed to the next-stage module providing the maximum amount of information. The provisional ability estimate was computed at the end of each stage using EAP estimation, as in the CAT designs. This routing method was used by Kim and Plake (1993) and is similar to the approximate maximum information (AMI) procedure used in several existing MST applications (e.g. Luecht et al., 2006).

MST Assembly

While several automated test assembly (ATA) software programs, such as CASTISEL (Luecht, 1998), were available for assembling MST modules and panels, none of them have been designed to do so under the relatively new TRT measurement model. Thus, instead of modifying existing ATA programs to work with the 3PL-TRT,

the panels and associated modules in this dissertation were assembled manually, as was done by Davis and Dodd (2003).

Test Construction Targets and Constraints

The task of constructing an MST involved formulating statistical targets that the modules and panels were built to and identifying any additional constraints that needed to be simultaneously satisfied. Additional constraints for MSTs typically arise from requirements in the test blueprint, content specifications, or to control the exposure rates for modules and items. In this dissertation, the statistical targets were defined as target test information functions (TIFs). This was similar to the approach taken by Luecht, Brumfield and Breithaupt (2006) to build MSTs with the 1-3-3 stage structure for an operational certification examination.

Following Luecht et al.'s (2006) example, TIFs were specified for the seven modules needed in the MST design for this study. The TIFs for the easy modules (2E and 3E) peaked at $\theta = -1$, the TIFs for the hard modules (2H and 3H) peaked at $\theta = 1$, and the TIFs for the three moderate module (1M, 2M, 3M) peaked at $\theta = 0$.

Three additional constraints related to the testlets used in the MST modules and panels were specified. These constraints were needed to maintain a similar maximum testlet and item exposure rates as the two CAT designs. They included,

1. No testlet may be used in more than two modules
2. No module may be part of more than two panels
3. If a testlet was used in two modules, the two modules may not be part of the same panel. This constraint prevented a testlet from appearing twice in any given path within a panel

Sub-pool Formation

Note that in this dissertation, four separate eight-panel MSTs needed to be assembled for each of the test length × item pool size conditions. The test length dictated the size of the testlets used in the modules. The long test length conditions required 10-item or 12-item testlets while the short test length required 5-item or 6-item testlets. The item pool size condition affected the number of testlets in the item pool available for building modules. Under the full item pool size condition, all 46 testlets were available while with the reduced pool size condition, only 31 testlets were available.

For each of the four test length × item pool size conditions, six MST *sub-pools* were formed from the available item pool. Each sub-pool was a scaled down version of the original pool. It had the same number of testlets as the original pool. However, the number of items associated with each testlet was restricted to those required by the test length condition. For example, each sub-pool formed for the long test length × full item pool condition still consisted of 46 testlets, but the number of items associated with each testlet was paired down to either 10 items or 12 items.

Six separate sub-pools were needed because within each MST panel, six types of modules were being assembled. For example, the long test length × full item pool condition required an easy 12-item testlet for Module 2E, an easy 10-item module for Module 3E 10-item, a hard 12-item testlet for Module 2H, a hard 10-item testlet for Module 3H, a moderate 12-item module for testlet 2M and three moderate 10-item modules for Modules 1M and 3M (see Figure 3.5). Correspondingly, sub-pools consisting of easy 12-item testlets, easy 10-item testlets, hard 12-item testlets, hard 10-item testlets, moderate 12-item testlets and moderate 10-item testlets were formed. This implied that six permutations of each testlet in the original pool were needed, one for each sub-pool. This requirement was not too different from what was needed for the

testlet-level CAT item pool. However, instead of building permutations that were parallel forms, the six permutations were built so that their testlet information functions resembled the shape of one of the TIFs in Figure 3.6. So, for example, easy 12-item permutations of each testlet were built to match the TIF of Module 2E and formed the easy 12-item testlet sub-pool; easy 10-item permutations of each built to match the TIF for Module 3E formed the easy 10-item testlet sub-pool, and so forth. This sub-pool formulation process was repeated for each of the four test length × item pool size conditions.

Module and Panel Assembly

With the sub-pools in place, the modules and panels were ready to be assembled. To build modules and panels, most ATA programs would define objective functions equal to the sum of the differences between a potential set of MST modules and their corresponding TIFs. Mathematical techniques such as liner programming (van der Linden, 1998) or optimization heuristics (Luecht & Nungester, 1998) were then applied to minimize the objective function while simultaneously satisfying additional constraints. Manual MST construction lacked the computational sophistication and efficiency afforded by ATA programs. However, it can emulate the process by judiciously choosing module candidates from each sub-pool, placing them into the MST stage structure, carefully comparing the modules for each panel within stage and across stages, and making any necessary adjustments.

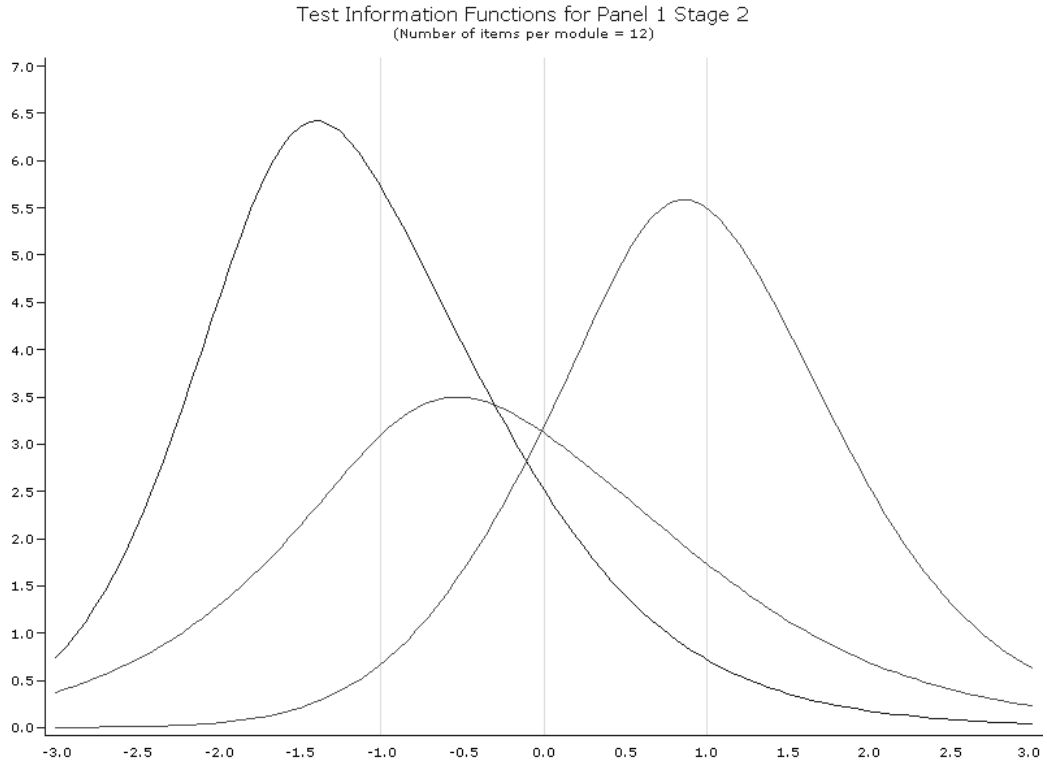

Figure 3.6: Information functions for one stage of an initially constructed panel

This was an iterative and time-consuming process. For this dissertation, the initial eight-panel MSTs constructed were deemed to be unacceptable after examining the test information functions of the modules within each panel. Many of the moderate and hard modules had substantially less information than the easy modules. This was an artifact of how test information was distributed in the overall item pool (see Figure 3.2). In addition, the test information functions of consecutive modules within each stage were often not spread out far enough on the θ scale. Figure 3.6 gives an example that illustrates these issues.

In this example, modules in the second stage (2E, 2M and 2H, respectively) of this panel had test information functions that peaked around the modes of their corresponding TIFs. However, disparity in the amount of information in each module

caused the information function of Module 2M to be almost completely overlapped by those of its neighboring modules (2E and 2H). The consequence of this within-stage unbalance of test information distribution was likely to be that very few examinees would be routed from Module 1M to Module 2M in this panel. Davis and Dodd (2003) encountered a similar issue in their manually assembled MST panel, which led to disproportional usage of the various paths an examinee could take through each panel.

To avoid the same pitfall, a second round of manual test assembly was conducted. Special emphasis was placed on putting together within-stage modules that had test information functions with similar heights and that were more evenly spread out along the θ scale. This process yielded panels with within-stage modules that were more balanced in test information. Figure 3.7 gives an example of the test information functions of one of the panel constructed for the long test × full pool size condition.

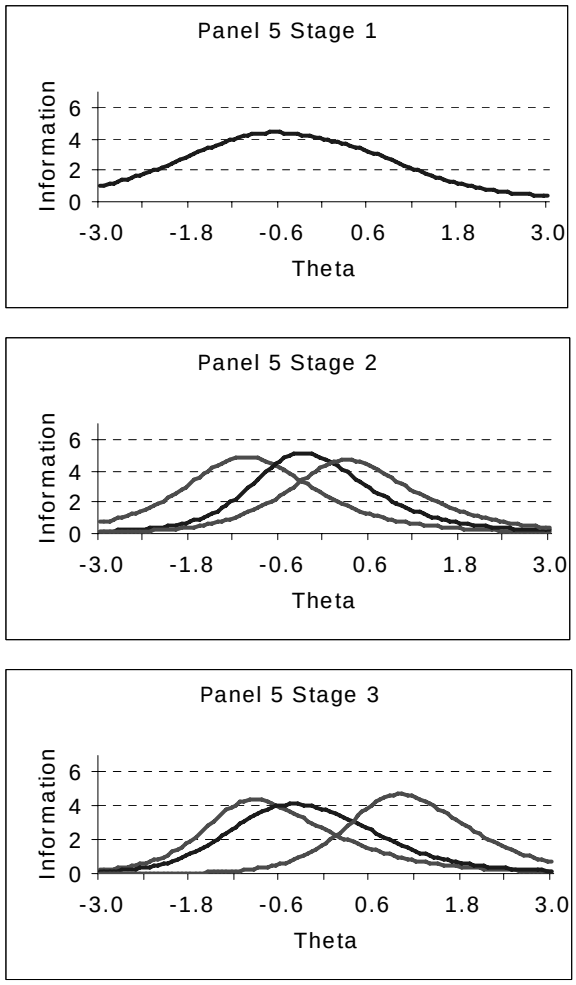

Figure 3.7: Test information functions for one of the panels

Figure 3.8 shows the test information functions for all first stage (1M) modules across the eight panels in the long test × full pool size condition. It shows that the test information function for the corresponding module across panels were generally similar.

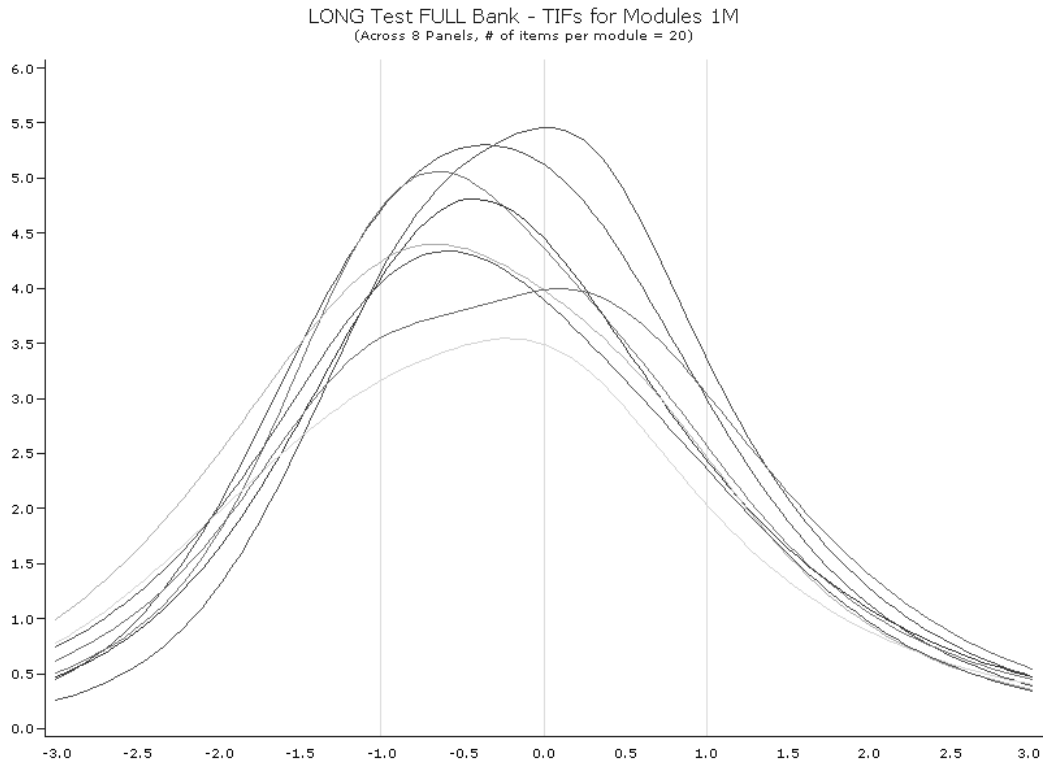

Figure 3.8: Test information functions for Module 1M across panels

Graphs similar to those shown in Figure 3.7 (within panel) and Figure 3.8 (across panels) were produced for every panel built across the four test length × item pool size conditions. Each graph was examined to ensure that the manual MST assembly process was achieving its objectives.

An additional consideration during the MST assembly process was to make use of as many unique testlets in the original item pool as possible in building the modules. Ideally, all available testlets would be part of at least one module in one of the eight panels. This would allow for more direct comparison of testlet exposure and pool utilization rates between the MST and CAT designs since all testlets in the pool are available for selection by the testing algorithm in the CAT designs. However, because of all the other targets and constraints that had to be satisfied in the MST assembly process,

it was very difficult to out every available testlet in at least one module. Furthermore, even if a testlet were selected for inclusion into a module, only the *items* that were assigned to this testlet's permutation in the corresponding sub-pools would be part of the module. As such, many items in the original item pool had no chance of being administered in the MST design. In summary, after all MST panels were assembled, the long test length × full item pool size condition made use of 37 (of the 46) testlets and 428 (of the 1,008) item in the original pool. The long test length × reduced item pool size condition used 30 (of the 31) testlets, representing 375 (of the 741) items available in the original pool. The short test length × full item pool size condition included 43 (of the 46) testlets, and 253 (of the 1,008) item from the original pool. The short test length × reduced item pool size condition made use of all 31 available in the original pool, but just 201 (of the 741) items available because only 5 or 6 items were assigned to each permutation of a testlet. The fact that only a subset of the original item pool had a chance to be administered was taken into account when computing the exposure control statistics for each condition in the MST design.

Test Administration

After the panels and modules were assembled, the MSTs were administered according to these steps:

1. The examinee was randomly assigned one of the eight panels.
2. The examinee was administered the two testlets in the first-stage module (1M) of the assigned panel.
3. At the end of Stage 1, the provisional ability ($\hat{\theta}$) parameter for the examinee was estimated using EAP.

4. The examinee was routed to the second-stage module (2E, 2M or 2H) whose test information functions yielded maximum information at $\hat{\theta}$ and the testlet in the module was administered.
5. At the end of Stage 2, provisional ability ($\hat{\theta}$) and testlet effect ($\hat{\gamma}$) parameters for the examinee were estimated using EAP.
6. The examinee was routed to the third-stage module (3E, 3M or 3H) whose test information functions yielded maximum information at $\hat{\theta}$ and the testlet in the module was administered.
7. After Stage 3, the final ability and testlet effect parameters were estimated with the entire set of responses using EAP estimation.

DATA ANALYSIS

The goal of this dissertation was to compare measurement accuracy and precision and the exposure control properties of the three CAT and MST designs across several manipulated test conditions.

Measurement accuracy and precision were assessed by the degree to which each test design recovered the known examinee θ values. This included computing and comparing the mean and standard error of the final θ estimates, and the Pearson product-moment correlation between the estimated and known θ values for each replication and grand means calculated across the 10 replications. Several indices of measurement effectiveness used in CAT and MST studies were also used. They included bias, root mean squared error (RMSE), and average absolute difference (AAD). The formulas for these indices are given below:

$$Bias = \frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)}{n} \tag{18}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{19}$$

$$\text{AAD} = \frac{\sum_{i=1}^{n}\left|\hat{\theta}_i - \theta_i\right|}{n} \tag{20}$$

In these formulas, $\hat{\theta}_i$ and θ_i represents the final ability estimate and known ability of examinee *i* respectively, and *n* is the total number of simulated examinees in each condition. All statistics were averaged across the 10 replications for each of the 24 study conditions.

Conditional plots of the mean bias and grand mean standard error across 10 replications were constructed in order to assess differences in the three adaptive test designs across the ability scale. Conditional plots can visually illustrate the measurement accuracy and precision of the designs in various parts of the θ scale. The test designs might be similar in their overall measurement effectiveness (as indicated by the mean bias and standard errors), but could still differ in measurement effectiveness at different θ values. A conditional plot would help capture such differences.

To evaluate the exposure control properties, item and testlet exposure rates were computed. An item's or testlet's exposure rate was computed by dividing the number of examinees to which the item or testlet was administered by the total number of examinees within each condition. The frequency distribution, mean, standard deviation and maximum of the exposure rates were computed and summarized across conditions. The proportion of items and testlets within the pool never administered was also calculated as an indicator of pool utilization. Also, measures of item and testlet overlap – the average number of items and testlet shared by two examinees – were computed. This was done separately for examinees of similar and of different abilities. Adopting the definition

used by Boyd (2003), *similar* examinees were defined as examinees whose known θ values differ by one logit or fewer; while *different* examinees were those whose known θ values differ by more than one logit. In order to compute the exposure rates and overlap statistics, a record of the items and testlets administered to each examinee, known as the examinee's audit trail, was recorded and analyzed. As with the measurement effectiveness statistics, all exposure control measures were averaged across the 10 replications for each of the 24 study conditions.

CHAPTER FOUR: RESULTS

The results for the three adaptive test designs are presented in this chapter. The three designs were compared on measurement accuracy and precision, and these findings are presented first. They are followed by the exposure rates and test overlap statistics used to assess the exposure control properties of the three designs. Each adaptive test design was simulated across three manipulated test conditions – test length, item pool size and underlying ability distribution. To save space, only tables and figures that contrast notable results from particular conditions are included in this chapter. Tables and figures for any conditions omitted from this chapter are given in Appendices B and C. All results were averaged across the ten replications in each study condition.

MEASUREMENT ACCURACY AND PRECISION

Measurement accuracy and precision were evaluated by the degree to which each test design recovered the known examinee θ values. Dependent measures computed included the mean and standard error of the final θ estimates, the Pearson product-moment correlation between the estimated and known θ values, bias, root mean squared error (RMSE), and average absolute difference (AAD). In addition, conditional plots of the mean bias and grand mean standard error were constructed to assess performance differences in the three test designs across the θ scale. Overall results of measurement effectiveness are given first in the next section, followed by the results of manipulating test length, item pool size and underlying ability distribution.

Overall

Tables 4.1 and 4.2 give the overall measurement accuracy and precision statistics for the three adaptive test designs. The *long test length, full item pool size and normal*

ability distribution condition was chosen to represent the overall results because it closely resembled the real testing parameters of the statewide reading assessment from which the data were drawn. These results served as a baseline to which the other manipulated conditions were compared.

Table 4.1: Descriptive Statistics of the estimated θ - overall results (long test length, full item pool, normal ability distribution condition)

	Estimated θ[a]		
Test Design	Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	-0.006 (-0.041, 0.032)	0.367 (0.364, 0.371)	0.919 (0.914, 0.926)
Item-Level CAT	-0.004 (-0.034, 0.038)	0.292 (0.290, 0.294)	0.938 (0.931, 0.944)
MST	-0.002 (-0.040, 0.046)	0.329 (0.328, 0.332)	0.929 (0.920, 0.937)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = -0.003, min mean = -0.047, max mean = 0.041

[b] SE: standard error; Correlation: between known and estimated θ's

These tables show that the overall measurement accuracy of the three adaptive test designs were good and very similar. The correlations between the estimated and known θ's were all above .90, with the correlation for the item-level CAT being the highest at .94. The biases for all three test designs were essentially zero, when rounded to the second decimal place. The AADs were also similar between the three test designs, with the AAD for the item-level CAT being the lowest at .27. The measurement accuracy results for the testlet-level CAT design were strikingly similar to those found for the analogous condition in Boyd's (2003) study. This was encouraging because the

testlet-level CAT was a direct extension of Boyd's (2003) progressive-restrictive (maximum exposure rate = .30) condition under the TRT model. Thus, it verified that the testlet-level CAT simulation was performing as expected. It also provided cross-validation to Boyd's (2003) finding on a completely different dataset

Table 4.2: Bias, RMSE and AAD of the estimated θ - overall results (long test length, full item pool, normal ability distribution condition)

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.004 (-0.020, 0.021)	0.394 (0.384, 0.404)	0.312 (0.300, 0.322)
Item-Level CAT	0.001 (-0.023, 0.019)	0.345 (0.331, 0.363)	0.273 (0.260, 0.288)
MST	-0.001 (-0.016, 0.016)	0.370 (0.344, 0.381)	0.293 (0.277, 0.300)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

Tables 4.1 and 4.2 also show that the overall measurement precision of the three test designs were slightly different. The item-level CAT most precisely measured θ, as it had the lowest mean standard error (SE) and lowest RMSE. The MST was second, followed by the testlet-level CAT design. The difference in overall measurement precision, however, was small and likely not practically significant. The difference in mean SEs, for example, between the item-level and test-level CAT was only about .08. The measurement precision results for the testlet-level CAT design were also similar to those found for the analogous condition in Boyd's (2003) study.

Figures 4.1 and 4.2 give the conditional plots of the mean bias and grand mean standard error for the long test length, full item pool size and normal ability distribution condition. Figure 4.1 shows that all three test design performed well and similarly in

terms of measurement accuracy across the θ scale. The characteristic reverse S-shaped curve implied that the test designs recovered the known θ values most accurately near the center of the ability distribution and least accurately at the extremes.

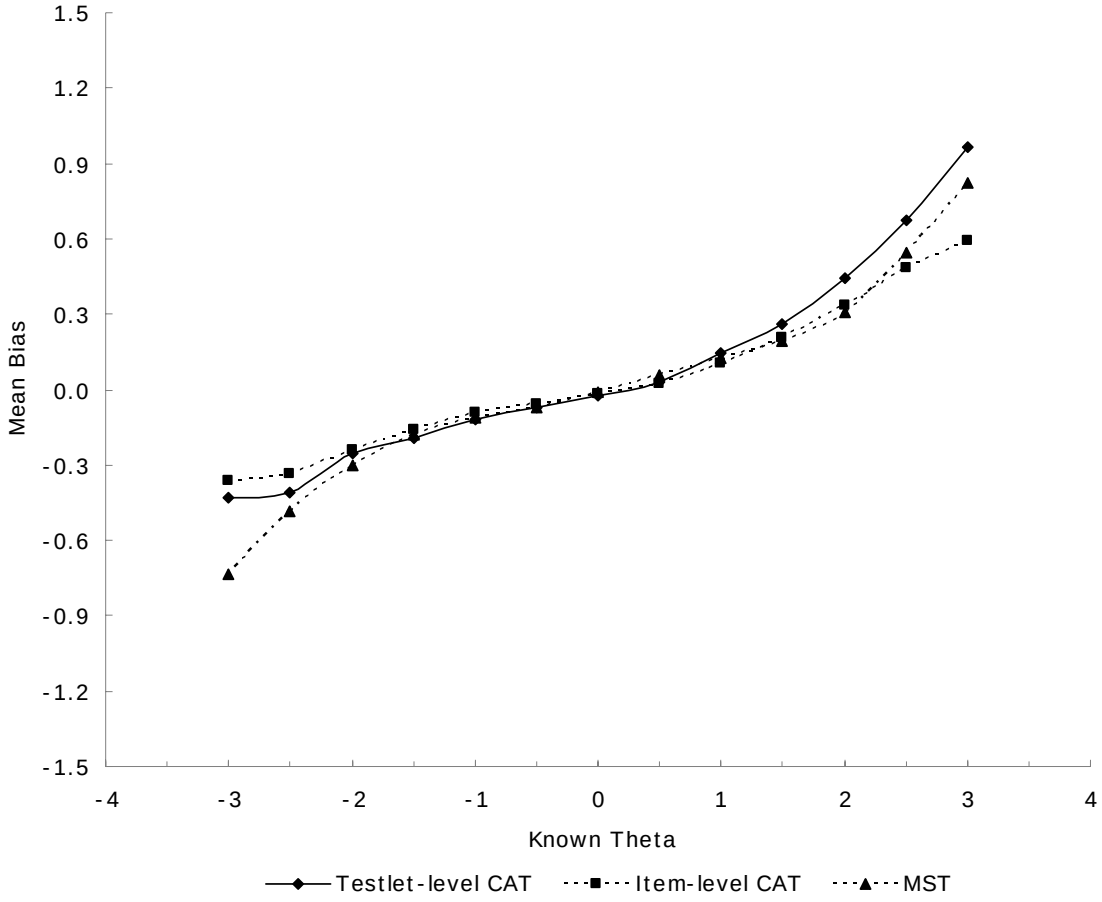

Figure 4.1: Conditional mean bias plot – overall results

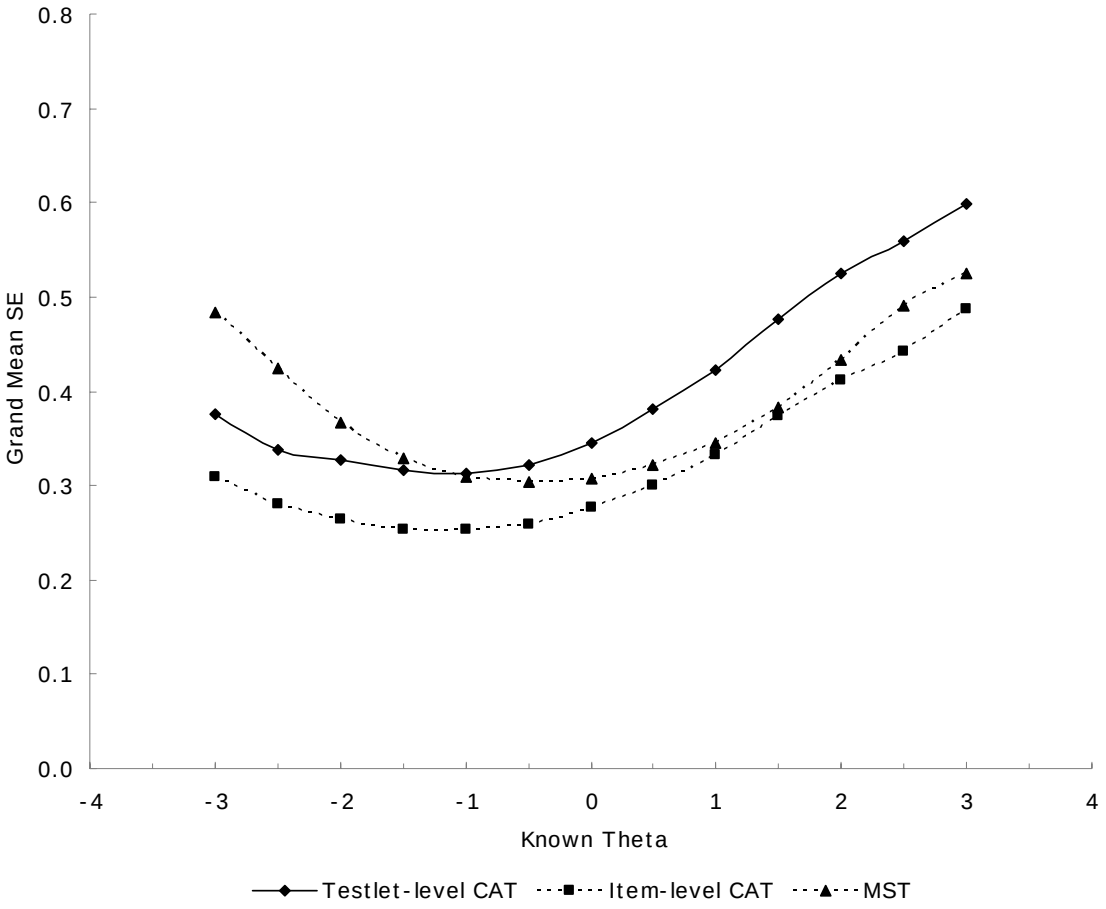

Figure 4.2: Conditional grand mean standard error plot – overall results

Figure 4.2, on the other hand, shows that the measurement precision across the θ scale was quite different for the three test designs. Each design had a U-shaped curve, which is characteristic of this type of conditional plot. However, the location and shape of the curves varied. The curves for the testlet-level and item-level CATs ran parallel to one another, with the curve for the item-level CAT being lower. Thus, the mean SE of the item-level CAT was consistently lower than that of the testlet-level CAT across the θ scale. And as seen from the overall descriptive statistics (Table 4.1), this consistent difference in grand mean SE was approximately .08. It should also be noted that the

vertical axis of symmetry of the U-shaped curve for each CAT condition was not around $\theta = 0$, but around $\theta = -1$. This was a direct reflection of the distribution of test information across the θ scale in the full item pool (see Figure 3.1).

The conditional grand mean standard error curve for the MST design did not parallel the other two curves. It was above the curve for the item-level CAT, but the distance between the curves was greater on the lower end of the θ scale. At the higher end of the scale, the difference between the two curves was practically negligible. This implied that the MST design was similar to the item-level CAT in measurement precision for examinees with high abilities, but was considerably less precise at measuring low-proficiency examinees. In relation to the testlet-level CAT, the curve for the MST design was below the curve for the testlet-level CAT at the high end of the θ scale, implying better precision for MST in measuring high-proficiency examinees. However, the curves actually crossed near $\theta = -1$ and the curve for the MST design was higher at the lower end of the θ scale. This meant that the MST design was less precise at measuring examinees with lower proficiencies when compared to the testlet-level CAT. Note also that the vertical axis of symmetry for the MST curve was around zero. This was a direct reflection of the distribution of test information across the θ scale in the MST panels (see Figure 3.7), which were built to satisfy the TIFs for the modules in each panel.

In summary, measurement accuracy of the three adaptive test designs was found to be very similar, both overall and across different points on the θ scale. Overall measurement precision of the three designs differed slightly, with the item-level CAT performing the best. Across the θ scale, however, the measurement precision of the MST design was considerably better for high θ values, especially when compared to the testlet-level CAT. But it was considerably worse than the item-level CAT for low θ values.

Test Length

Tables 4.3 and 4.4 give the measurement effectiveness statistics for the *short* test length, full item pool and normal ability distribution condition. The results in these tables were contrasted with those in Tables 4.1 and 4.2 to assess the effect of shortening the total test length from 42 items to 21 items.

These tables show that the overall measurement accuracy of the three adaptive test designs was still good with the shorter test, but differences were noticeable. For example, the correlations between the estimated and known θ's were still relatively high, ranging from .87 to .91, but these correlations were all lower compared to those in the longer test. The drop in correlation was also different for the three test designs. The item-level CAT only dropped about .02 while both the testlet-level CAT and the MST designs dropped .05 in their correlations. The bias values for all three test designs were still close to zero, but the AADs for the designs were higher compared to those of the longer test length, ranging from .32 to .39. The AAD for the item-level CAT was least affected by the change in test length, only going up by .05. The AADs for the testlet-level CAT and MST designs, on the other hand, each went up by .08. Thus, while shortening the test length did have a small effect on measurement accuracy, the effect seemed slightly greater on the testlet-level CAT and MST designs.

In terms of overall measurement precision, Tables 4.3 and 4.4 show that the item-level CAT still performed the best with the lowest mean SE and RMSE of .37 and .41 respectively. The MST design was again second, followed by the testlet-level CAT. Comparing these tables with Tables 4.1 and 4.2 showed that the overall measurement precision of all test designs dropped with the shorter test, not an unexpected result. Shortening the test, however, seemed to least affect the measurement precision of the

item-level CAT design. Its mean SE, for example, only went up by .08, while the mean SE for the testlet-level CAT and MST designs went up by .15 and .14 respectively.

Table 4.3: Descriptive Statistics of the estimated θ - short test length (short test length, full item pool, normal ability distribution condition)

	Estimated θ[a]		
Test Design	Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	-0.004 (-0.036, 0.046)	0.515 (0.511, 0.522)	0.870 (0.852, 0.879)
Item-Level CAT	0.000 (-0.028, 0.037)	0.367 (0.363, 0.369)	0.914 (0.908, 0.921)
MST	0.004 (-0.047, 0.049)	0.468 (0.465, 0.470)	0.880 (0.873, 0.885)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = -0.003, min mean = -0.047, max mean = 0.041

[b] SE: standard error; Correlation: between known and estimated θ's

Table 4.4: Bias, RMSE and AAD of the estimated θ - short test length (short test length, full item pool, normal ability distribution condition)

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.001 (-0.034, 0.032)	0.493 (0.477, 0.509)	0.389 (0.378, 0.401)
Item-Level CAT	-0.003 (-0.023, 0.008)	0.405 (0.387, 0.419)	0.318 (0.307, 0.328)
MST	-0.007 (-0.018, 0.005)	0.473 (0.46, 0.492)	0.374 (0.363, 0.387)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

Figures 4.3 and 4.4 give the conditional mean bias and grand mean standard error plots for the *short* test length, full item pool size and normal ability distribution condition. The general shapes of the curves in these figures were similar to those observed for the long test length (Figures 4.1 and 4.2). Thus, similar conclusions may be drawn about the measurement characteristics of the three designs across the θ scale for the short test length. A few key differences should be noted though.

First, in contrast to Figure 4.1, Figure 4.3 showed higher mean biases for all test designs at the two ends of the θ scale. The differences in mean biases between the three designs were also more pronounced at the ends of the θ scale. These implied that the slight decrease in measurement accuracy observed in the overall statistics (Tables 4.3 and 4.4) was mainly attributable to the measurement of examinees with very high or very low known abilities. Also, shortening the test had a differential effect on measurement accuracy at the ends of the θ scale, with the effect being smallest for the item-level CAT.

Next, in contrast to Figure 4.2, Figure 4.4 also showed higher grand mean standard errors for all three test designs, but here it was observed across the entire θ scale. In addition, the curves for the testlet-level CAT and MST designs are both notably higher than the curve for the item-level CAT. These implied that the measurement precision for all three designs decreased when the test was shortened, and this effect was found for all known θ values, not just the extremes ones. Also, the differential effect of test length on measurement precision was found across the θ scale, with the effect again being smallest for the item-level CAT.

In summary, the results showed that shortening the test length had the effect of decreasing both measurement accuracy and precision. The decrease in measurement accuracy was found to be most prevalent at the ends of the θ scale, while the decrease in

precision was observed consistently across the θ scale. The effect of shortening the test appeared to be smaller for the item-level CAT than it was for the other two test designs.

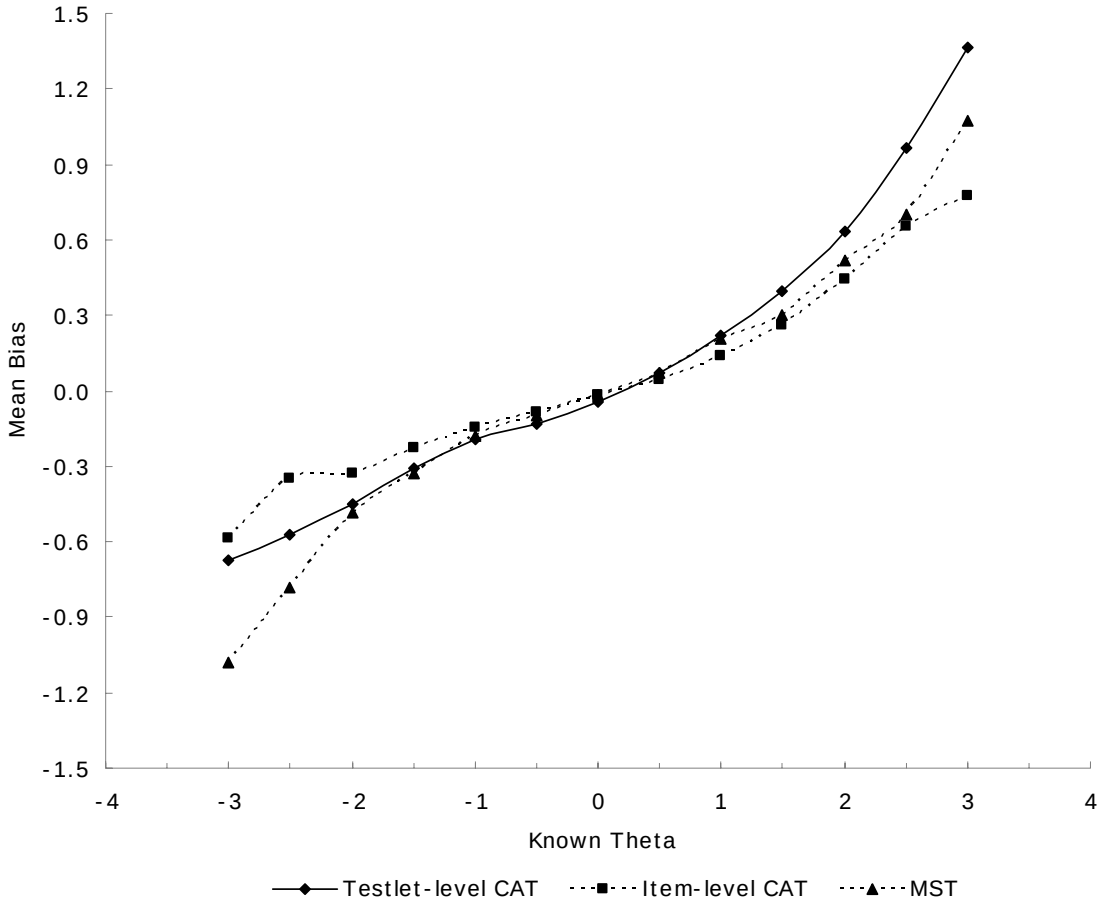

Figure 4.3: Conditional mean bias plot – short test length

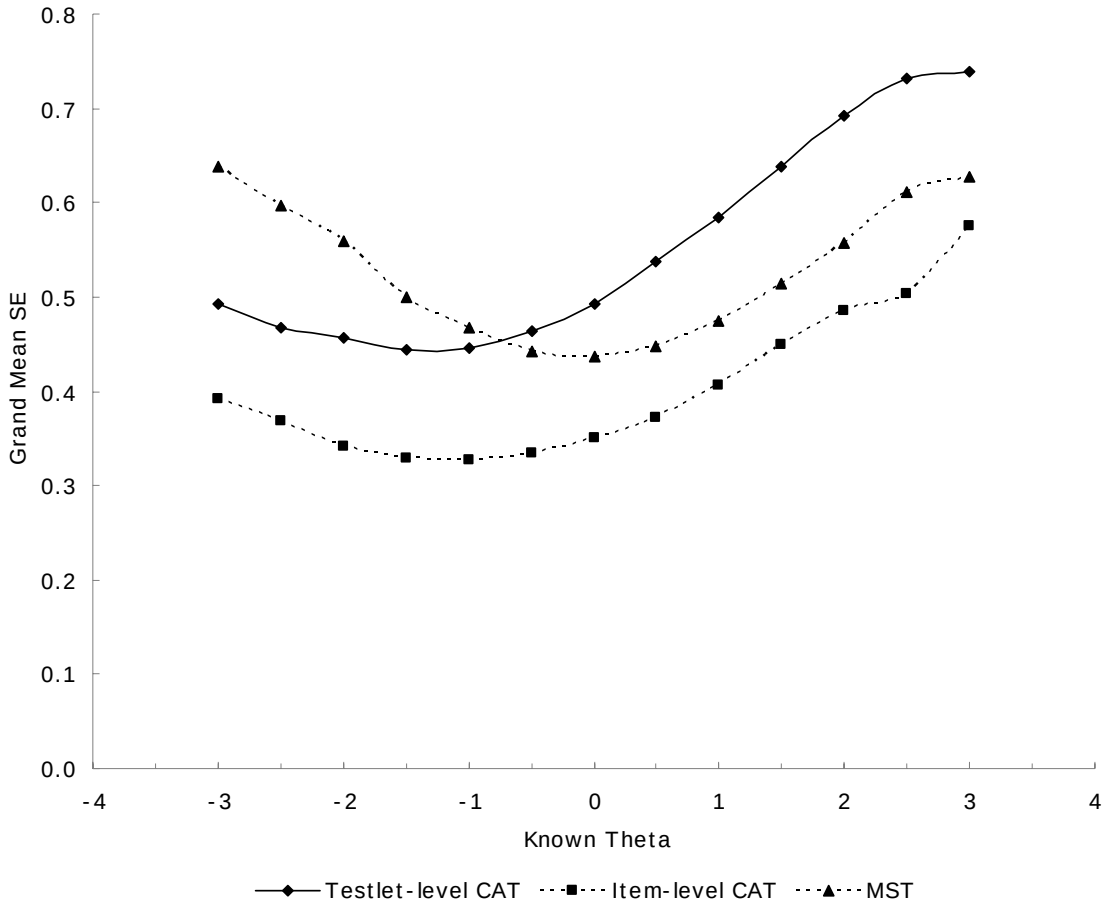

Figure 4.4: Conditional grand mean standard error plot – short test length

Pool Size

Tables 4.5 and 4.6 give the measurement accuracy and precision statistics for the long test length, *reduced* item pool and normal ability distribution condition. The reduced item pool had only 31 testlets and 741 items, roughly two-third of the size of the full item pool. The results in these tables were compared to those in Tables 4.1 and 4.2 to evaluate the effect of reducing the item pool size on measurement effectiveness.

Table 4.5: Descriptive Statistics of the estimated θ - reduced pool (long test length, reduced item pool, normal ability distribution condition)

	Estimated θ[a]		
Test Design	Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	0.002 (-0.036, 0.051)	0.368 (0.363, 0.371)	0.915 (0.904, 0.924)
Item-Level CAT	-0.004 (-0.046, 0.047)	0.291 (0.289, 0.294)	0.941 (0.936, 0.948)
MST	-0.001 (-0.031, 0.052)	0.328 (0.326, 0.331)	0.929 (0.925, 0.936)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = -0.003, min mean = -0.047, max mean = 0.041

[b] SE: standard error; Correlation: between known and estimated θ's

Table 4.6: Bias, RMSE and AAD of the estimated θ - reduced pool (long test length, reduced item pool, normal ability distribution condition)

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	-0.005 (-0.015, 0.007)	0.403 (0.38, 0.424)	0.319 (0.301, 0.334)
Item-Level CAT	0.001 (-0.015, 0.013)	0.338 (0.33, 0.347)	0.266 (0.259, 0.274)
MST	-0.002 (-0.016, 0.013)	0.369 (0.348, 0.381)	0.292 (0.283, 0.300)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

The results in Tables 4.5 and 4.6 were strikingly similar to those in Tables 4.1 and 4.2. The mean correlation values, for example, were .92, .94 and .93 respectively for the testlet-level CAT, item-level CAT and MST designs in the reduced item pool condition. These values were identical (when rounded to the second decimal place) to the corresponding ones for the full item pool condition. The RMSEs for the three test designs were .40, .35 and .37 respectively in the reduced item pool, again identical (when rounded to the second decimal place) to the corresponding values in the full item pool. These results implied that reducing the pool size did not appear to have any discernable impact on the measurement properties for any of the three test designs.

Figures 4.5 and 4.6 give the conditional mean bias and grand mean standard error plots for the long test length, *reduced* item pool and normal ability distribution condition. As with the overall measurement statistics, these plots look virtually identical to corresponding ones for the full item pool condition (Figures 4.1 and 4.2). Thus, across

the θ scale, reducing the pool size appeared to have no discernable impact on the measurement accuracy and precision of the three test designs.

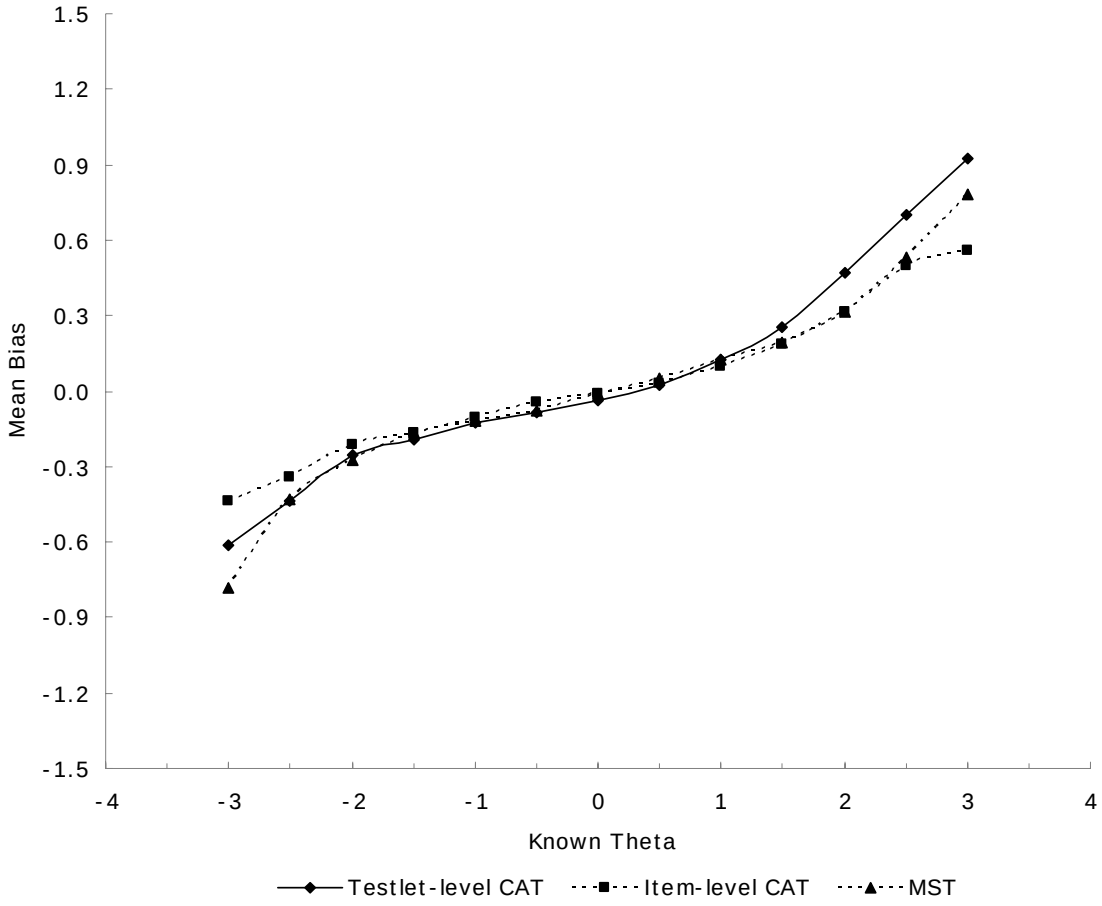

Figure 4.5: Conditional mean bias plot – reduced pool

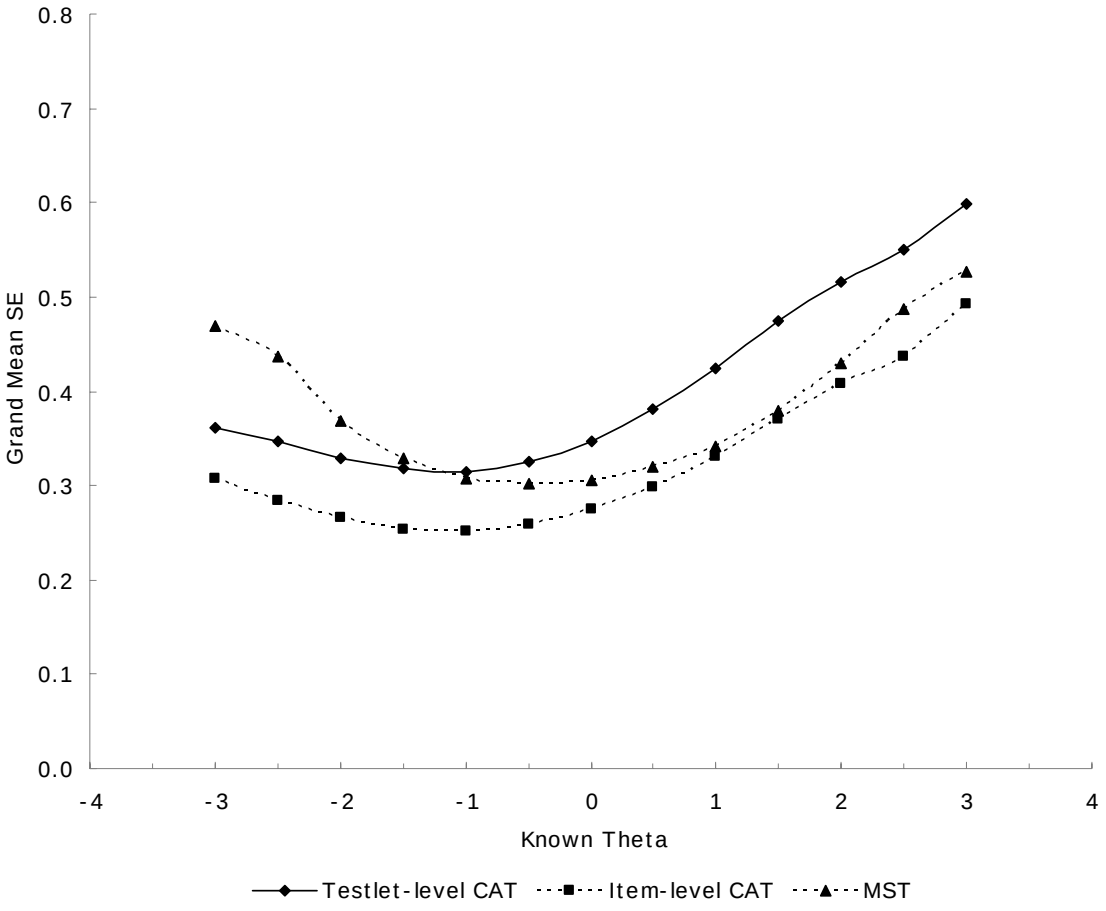

Figure 4.6: Conditional grand mean standard error plot – reduced pool

Test Length × Pool Size Interaction

Recall that the reduced item pool size was chosen to be two-thirds (instead of one-half) the size of the full item pool so that different combinations of item-pool-to-test length ratios would result across the study conditions (see Table 3.1). The purpose of this was to see if a two-way test length × pool size interaction existed for any of the test designs. To complete the check for interaction effects, the measurement effectiveness statistics and conditional plots for the *short* test length, *reduced* item pool and normal distribution condition (in Tables and Figures 4.7 and 4.8) were compared to those for the *short* test length, *full* item pool and normal distribution condition (in Tables and Figures 4.3 and 4.4). As with the long test length, the corresponding tables and figures between these two conditions were nearly identical. Thus, reducing the pool size with the shorter test also had no discernable effect on the measurement properties of the three test designs. Thus, no test length × pool size interaction effect seemed to be present.

In summary, the results showed that reducing the size of the item pool had no notable impact on the measurement effectiveness of any of three test designs. This was observed for both the long and short test lengths. As such, no test length × pool size interaction effect appeared to exist either.

Table 4.7: Descriptive statistics of the estimated θ (short test, reduced pool, normal ability distribution condition)

Test Design	Estimated θ[a] Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	-0.004 (-0.051, 0.051)	0.516 (0.512, 0.523)	0.871 (0.859, 0.88)
Item-Level CAT	-0.004 (-0.057, 0.076)	0.366 (0.364, 0.370)	0.918 (0.906, 0.923)
MST	0.011 (-0.043, 0.064)	0.467 (0.463, 0.470)	0.879 (0.864, 0.885)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = -0.003, min mean = -0.047, max mean = 0.041

[b] SE: standard error; Correlation: between known and estimated θ's

Table 4.8: Bias, RMSE and AAD of the estimated θ (short test, reduced pool, normal ability distribution condition)

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.001 (-0.009, 0.016)	0.491 (0.477, 0.503)	0.388 (0.375, 0.401)
Item-Level CAT	0.001 (-0.035, 0.029)	0.397 (0.382, 0.419)	0.313 (0.302, 0.329)
MST	-0.014 (-0.028, -0.001)	0.477 (0.464, 0.490)	0.376 (0.366, 0.387)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

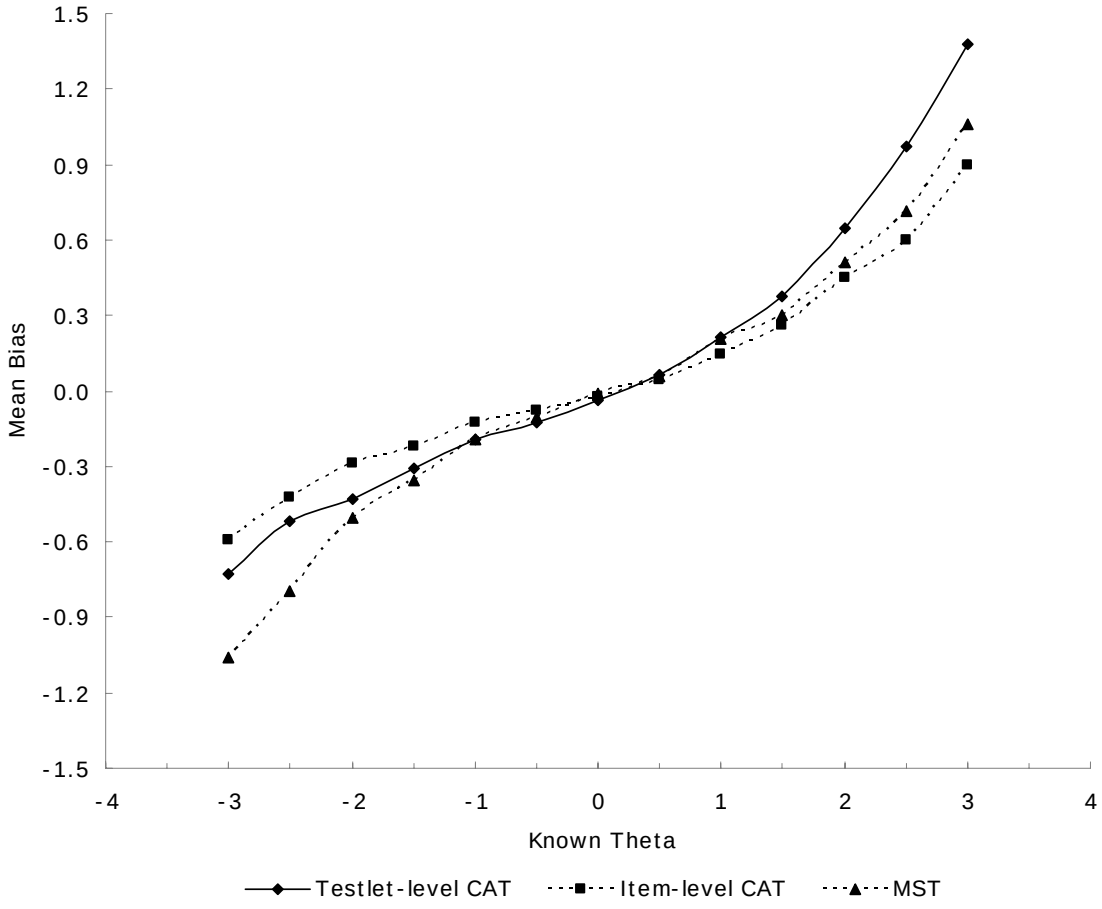

Figure 4.7: Conditional mean bias plot – short test, reduced pool

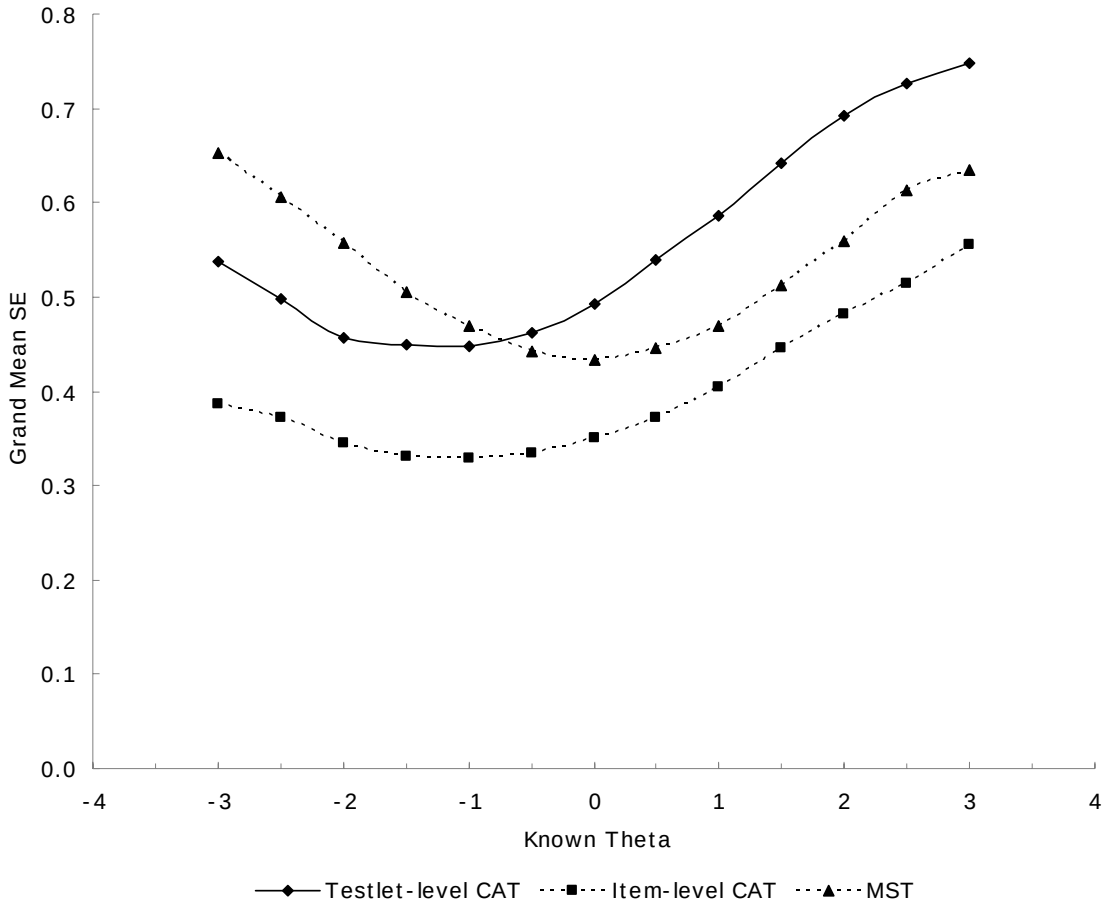

Figure 4.8: Conditional grand mean standard error plot – short test, reduced pool

Ability Distribution

Tables 4.9 and 4.10 give the measurement effectiveness statistics for the long test length, full item pool, and *skewed* ability distribution condition. For this condition, instead of a standard normal distribution, the examinee θ values were generated from a beta distribution with $\alpha = 5.0$ and $\beta = 1.8$ and transformed to be centered at zero with a standard deviation of one. This yielded a negatively-skewed θ distribution whose mean was at $\theta = 1.5$ (see Figure 3.3). The results from this condition were compared to those of the long test length, full item pool, and *normal* distribution condition in Tables 4.1 and 4.2 to assess the effect of varying the underlying ability distribution.

Table 4.9: Descriptive statistics of the estimated θ - skewed distribution (long test length, full item pool, skewed ability distribution condition)

	Estimated θ[a]		
Test Design	Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	1.130 (1.081, 1.163)	0.485 (0.477, 0.490)	0.883 (0.874, 0.893)
Item-Level CAT	1.222 (1.173, 1.262)	0.388 (0.383, 0.393)	0.912 (0.903, 0.920)
MST	1.213 (1.169, 1.263)	0.406 (0.401, 0.412)	0.914 (0.905, 0.920)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = 1.497, min mean = 1.426, max mean = 1.548

[b] SE: standard error; Correlation: between known and estimated θ's

Table 4.10: Bias, RMSE and AAD of the estimated θ - skewed distribution (long test length, full item pool, skewed ability distribution condition)

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.367 (0.337, 0.398)	0.600 (0.573, 0.618)	0.486 (0.46, 0.504)
Item-Level CAT	0.276 (0.253, 0.292)	0.497 (0.475, 0.513)	0.398 (0.381, 0.412)
MST	0.284 (0.243, 0.306)	0.500 (0.471, 0.523)	0.405 (0.379, 0.423)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

From Tables 4.9 and 4.10, it can be seen that the overall measurement accuracy was quite similar for the three test designs, although slightly worse for the testlet-level CAT. The mean correlations were nearly identical (.91) for the item-level CAT and MST, but it was slightly lower for the testlet-level CAT (.88). This was also the case for the bias values – the item-level CAT and MST designs had biases of .28 while the testlet-level CAT had a bias of .37. Compared to the results for the normal distribution condition (Tables 4.1 and 4.2), the mean correlations for the skewed distribution condition were marginally lower. The biases, however, were substantially higher compared to normal distribution, where the estimates were essentially unbiased. Thus, the skewed ability distribution seemed to cause all three designs to produce mean overall estimated θ values that were substantially lower than the mean of the known θ values. This effect seemed to be greater for the testlet-level CAT than it was for the other two test designs.

In terms of overall measurement precision, the mean standard error and RMSE values in Tables 4.9 and 4.10 show that item-level CAT and MST conditions were similar

while the testlet-level CAT was relatively worse. For example, both the item-level CAT and MST design has RMSE values of about .50 while the RMSE for the testlet-level CAT was .60. Compared to the results from the normal distribution condition, the measurement precision was worse for all three test designs. The mean standard error was higher by .12, .10 and .08 respectively for the testlet-level CAT, item-level CAT and MST design. The MST design, however, seemed slightly less affected by the change in underlying θ distribution, especially when compared to the testlet-level CAT.

Figures 4.9 and 4.10 give the conditional plots for the mean bias and grand mean standard error across the θ scale. Because there were very few examinees with known θ values less than -2 in the skewed distribution, those portions of the curves were truncated. Consequently, the characteristic reversed S-shaped and U-shaped curves were not as apparent in these two figures. The plots, however, provide a plausible explanation for what was observed in the overall measurement statistics. Figure 3.3 showed that a high proportion of the examinees in the skewed distribution had θ value greater than 2. Figures 4.9 and 4.10 show that this was the same range in which the θ estimates were the least accurate and precise. Thus, the higher proportion of poorly-estimated θ values caused the overall mean accuracy (bias) and precision (standard error) to be worse than when the underlying θ values were normally distributed. This effect was slightly smaller for the MST design because the θ values whose proportions had significantly decreased (i.e. θ values < -2) corresponded to the part of the θ scale for which the MST design had the highest mean standard errors, relative to the two CAT designs.

In summary, the results found that changing the underlying ability distribution from normal to negatively-skewed decreased the overall measurement accuracy and precision of all three test designs. This was due in large part to the higher proportion of θ

values in the high range of the ability scales that were poorly estimated. The decrease in measurement precision, however, was smaller for the MST design.

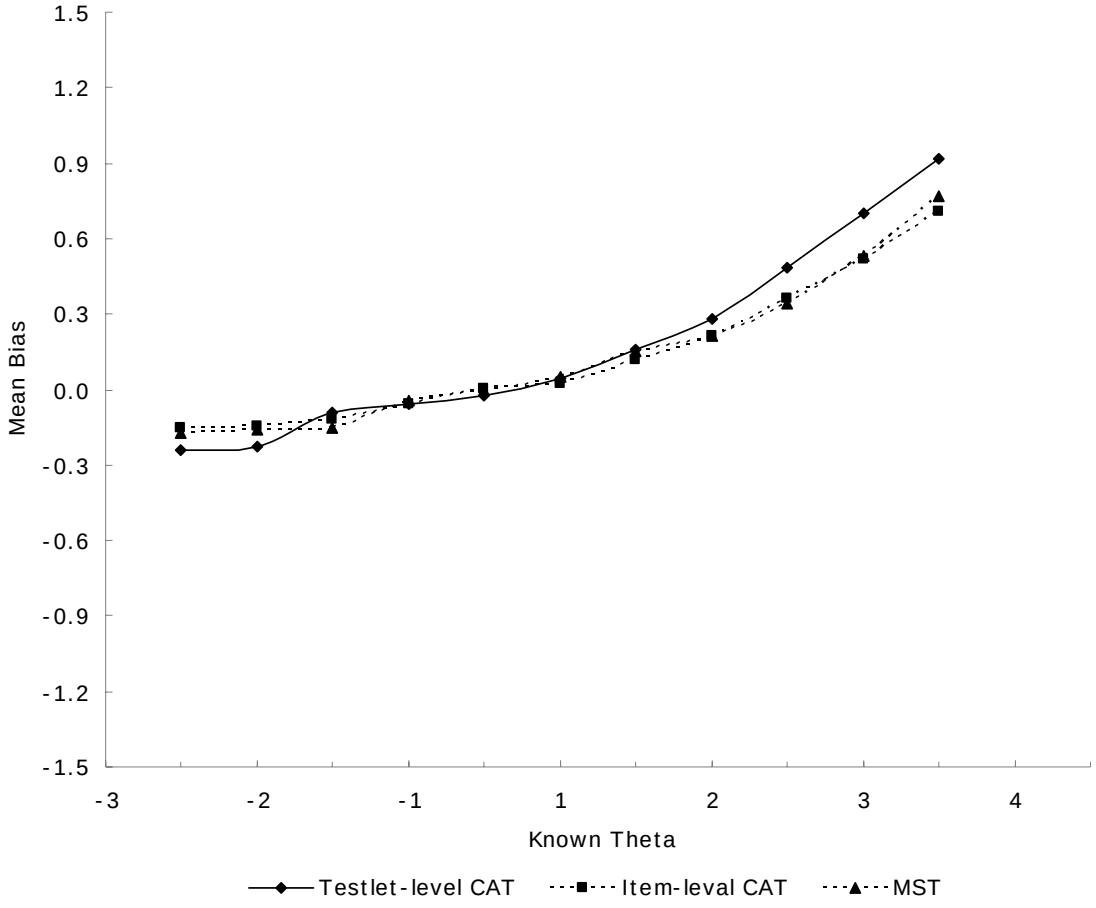

Figure 4.9: Conditional mean bias plot – skewed distribution

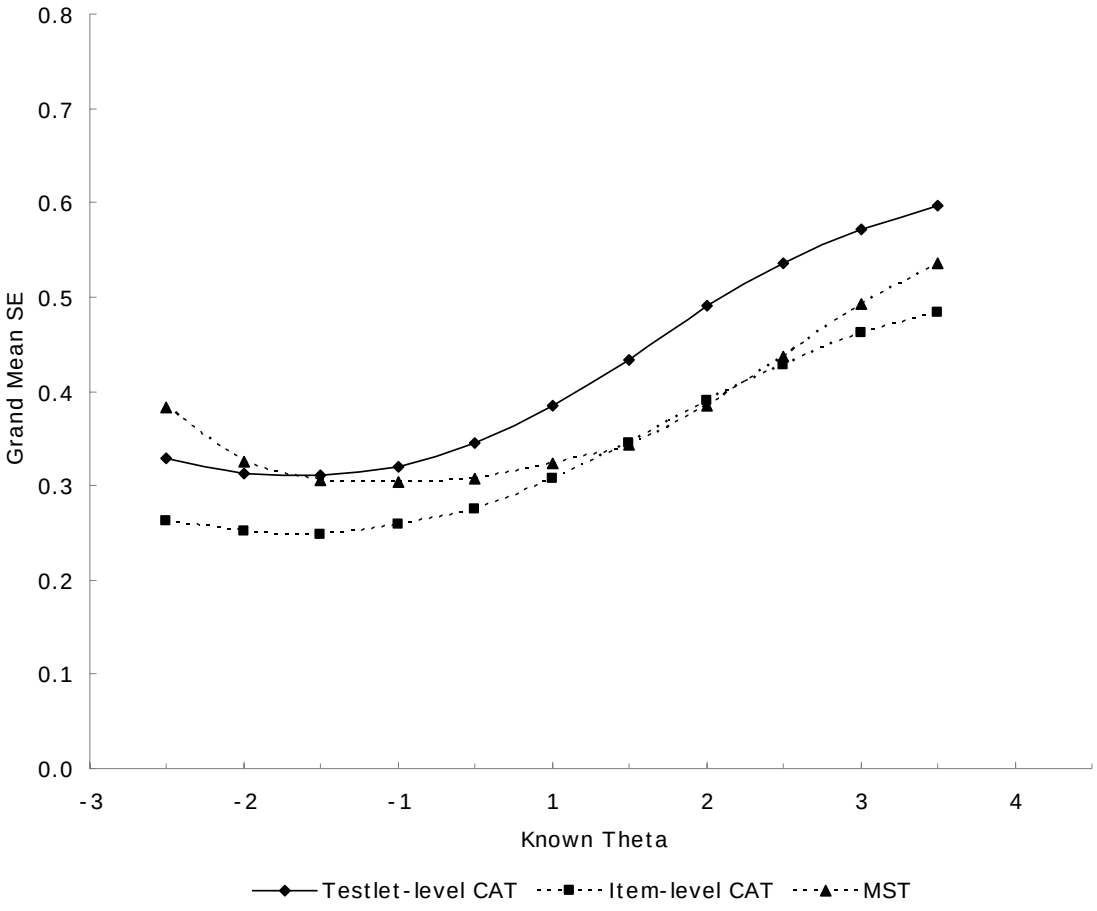

Figure 4.10: Conditional grand mean standard error plot – skewed distribution

EXPOSURE CONTROL PROPERTIES

Exposure control properties were assessed by computing descriptive statistics and frequency distribution of the item and testlet exposure rates. The average percentages of testlets and item never administered to any examinee over the ten replications were used as indicators of pool utilization. Item and testlet overlap rates were also calculated. This was done for all examinees as well as separately for examinees of similar and of different abilities. Similar examinees were defined as examinees whose known θ values differ by one logit or fewer, while different examinees were those whose known θ values differ by more than one logit (Boyd, 2003). Overall exposure control findings are provided in the next section, followed by the results from manipulating test length, pool size and ability distribution.

Overall

As with the measurement effective results, the *long test length, full item pool and normal distribution* condition was used to assess and compare the overall exposure control properties of the three test designs.

Testlet Exposure Rates

Table 4.11 gives the descriptive statistics for the testlet exposure rates. The results for the CAT designs were virtually identical. This came as no surprise because the between-testlet CAT algorithms were essentially the same for the two CAT designs. Note also that the mean maximum testlet exposure rates of the CAT designs were both .30, implying that the testlet-level progressive restricted procedure implemented in the CAT algorithm was successful at enforcing its maximum exposure rate.

The results for the MST looked slightly different. However, this was because not all 46 testlets were used in the eight assembled MST panels. The grand mean of the

testlet exposure rates was an indicator of this fact. Chen, Ankenmann and Spray (2003) showed that the mean exposure rate was directly related to the ratio of test length to pool size. This was demonstrated here as the mean testlet exposure rates for the CAT (.087) and MST (.108) designs were inversely proportional to the number of available testlets in the two designs (46 for CAT and 37 for MST).

Table 4.12 gives the frequency distribution of the testlet exposure rates averaged across the ten replications. Again, the results for the two CAT designs were nearly identical. All 46 testlets in the pool, on average, were administered at least once, showing good pool utilization. A total of 34 testlets (or 74% of the available testlets) had exposure rates less than .10, indicating consistently low testlet exposure to examinees.

The MST design had substantially less testlets with low exposure rates. On average, only 16 of the 37 testlets (or 43%) had testlet exposure rates less than .10. However, all testlets were administered at least once by the MST design, showing good pool utilization. The mean maximum exposure rate was also lower (.26) than those of the CAT designs.

Thus, in summary, the three adaptive test designs performed similarly well in terms of testlet exposure control.

Table 4.11: Descriptive statistics of testlet exposure rates – overall (long test length, full item pool, normal ability distribution condition)

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.087	0.088	0.301
Item-Level CAT	0.087	0.087	0.301
MST[a]	0.108	0.063	0.262

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 37 (of 46) testlets in its assembled panels

[b] SD: Standard Deviation

Table 4.12: Frequency distribution of testlet exposure rates – overall (long test length, full item pool, normal ability distribution condition)

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	1	2	0
.26-.30	4	4	1
.21-.25	1	1	2
.16-.20	2	2	5
.11-.15	4	4	14
.06-.10	9	9	6
.01-.05	25	25	10
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	46	46	37

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Item Exposure Rates

Tables 4.13 and 4.14 give the descriptive statistics and frequency distribution for the item exposure rates. In these tables, differences in item selection procedures for the two CAT designs were quite apparent. The grand mean of the item exposure rates were the same because of equal test-length-to-pool-size ratios (Chen, Ankenmann & Spray, 2003). However, the testlet-level CAT performed better at item exposure control. The mean maximum item exposure rate was only .20, the lowest of the three test designs. All 1,008 items in the pool were, on average, administered at least once, indicating good pool utilization. Also, an average of 729 items (or 73% of the pool) was administered very sparingly with exposure rates between .01 and .05. And an additional 176 items (or 18% of the pool) had exposure rates between .06 and .10. Thus, over 90% of the pool was given at least once, but to less than 10% of the examinees, a desirable item exposure property, especially if test security is a concern.

The item-level CAT also had a substantial proportion of pool items with low exposure rates. An average of 633 items (63% of the pool) had exposure rates between .01 and .05, and an additional 119 items (12% of the pool) had exposure rates between .06 and .10. So, about 75% of the item pool was given at least once, but to less than 10% of the examinees. The main issue with the item-level CAT, however, was its poor pool utilization. On average, 139 items (or 14% of the item pool) were never administered to any examinee. However, the mean maximum item exposure rate of .25 implied that the item-level progressive-restrictive procedure implemented within the item-level CAT algorithm was successful at maintaining its specified maximum exposure rate.

The MST design yielded a similar mean maximum item exposure rates (.26) as the item-level CAT. The grand mean of the item exposure rate was over two times higher than that of the CAT designs. This, however, was due to the fact that the number

of items used in the MST panels (428 items) was less than half the size of the full item pool (1,008 items) available to the CAT designs (Chen, Ankenmann & Spray, 2003). On the positive side, all 428 item used in the MST panels were administered at least once, thus showing good pool utilization. However, the distribution of item exposure rates was bimodal, peaking at both the .11 to .15 range and the .01 to .05 range. Of particular concern is how, on average, 147 items (or 34% of the available items) had exposure rates between .11 and .15, and an additional 69 items (or 16% of the available items) had exposure rates over .16. This implied that about 50% of the items available for administration were administered to more than 11% of the examinees, a notable test security concern in practice if examinees share test items with one another.

In summary, the testlet-level CAT was the best design at controlling item exposure while making use of all items in the pool. The item-level CAT design was also able to keep most its item exposure rates low while maintaining a specified maximum exposure rate. However, an average of 14% of the item pool was never administered. The MST design, in contrast, had good pool utilization. The considerably higher percentage of items with high exposure rates though could be a test security concern.

Table 4.13: Descriptive statistics of item exposure rates – overall (long test length, full item pool, normal ability distribution condition)

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.042	0.037	0.204
Item-Level CAT	0.042	0.056	0.251
MST[a]	0.098	0.062	0.262

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 428 (of 1,008) items in its assembled panels

[b] SD: Standard Deviation

Table 4.14: Frequency distribution of item exposure rates – overall (long test length, full item pool, normal ability distribution condition)

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	0	0	1
.26-.30	0	4	7
.21-.25	1	40	23
.16-.20	13	23	38
.11-.15	89	51	147
.06-.10	176	119	63
.01-.05	729	633	149
Not Admin	0	139	0
Not Admin %	0%	14%	0%
Total Items	1008	1008	428

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Overlap Rates

Table 4.15 gives the testlet overlap rates for the three adaptive test designs. It shows that the three designs performed similarly well in minimizing the overlap of testlets between examinees. In general, examinees had less than one testlet in common on their tests. This was true of the average testlet overlap rate for all examinees (ranging from .6 - .7), as well as for examinees of similar (ranging from .5 to .6) and different abilities (ranging from .7 to .8).

Table 4.16 gives the item overlap rates for the three designs. The item overlap rates appeared to be more different between the three test designs than they were for the testlet overlap rates. However, one should remember that the length of the test in this condition was 42 items. A difference in overlap of one or two items (out of 42) between examinees was likely not of any practical significance. Thus, while, in general, the MST design had the highest overlap of items between examinees – similar, different or overall – and testlet-level CAT design had the lowest, these differences in item overlap rates were probably not practically significant.

So in summary, the three adaptive test designs performed similarly well in terms of testlet and item overlap rates.

Table 4.15: Testlet overlap rates – overall (long test length, full item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0
Item-Level CAT	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0
MST	0.6	0.0	4.0	0.5	0.0	4.0	0.7	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.16: Item overlap rates – overall (long test length, full item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	3.1	0.0	37.3	2.0	0.0	32.0	3.3	0.0	37.3
Item-Level CAT	4.9	0.0	39.6	2.0	0.0	29.8	5.4	0.0	39.6
MST	5.8	0.0	42.0	3.8	0.0	42.0	6.1	0.0	42.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Test length

Testlet Exposure Rates

Tables 4.17 and 4.18 give the descriptive statistics and frequency distribution of the testlet exposure rates for the *short* test length, full item pool and normal distribution condition. By comparing the results in these tables with the corresponding ones in Tables 4.11 and 4.12, the effect of shortening the test on testlet exposure rates can be assessed.

The results in these two tables were very similar to what was observed for the long test length condition. This was not unexpected because in the short test condition, it is the number of items that has decreased (42 to 21 items), not the number of testlets (still 4 testlets). As such, the exposure properties at the testlet level were barely affected, and all three test designs performed similarly well.

Item Exposure Rates

Tables 4.19 and 4.20 give the descriptive statistics and frequency distribution of the item exposure rates for the *short* test length, full item pool and normal distribution condition. Results in these tables are contrasted with those in Tables 4.13 and 4.14 to evaluate the effect of shortening the test on item exposure rates.

In general, shortening the test length decreased item exposure rates for all three designs. This is not surprising because when the test-length-to-pool-size ratio decreases, the mean item exposure rate is expected to decrease as well (Chen, Ankenmann and Spray, 2003). This was observed in the CAT designs: the test-length-to-pool-size ratio decreased from 1:24 to 1:48, and consequently, the mean item exposure rate in this condition was half (.021) of what it was in the long test length condition (.042). A similar proportional decrease in the mean item exposure rate was also found for the MST design.

There were, however, differential effects on the three test design. For the testlet-level CAT, shortening the test substantially decreased the mean maximum item exposure rate (down to .10 from .20). As indicated by the mean standard deviation and the frequency distribution, the distribution of exposure rates across items was very compact, with most item having very low exposure rates (less than .05). Yet, at the same time, all but one item were administered to at least one examinee on average. This represented the ideal scenario in terms of controlling exposure while maintaining good pool utilization.

For the item-level CAT and MST designs, downward shifts in the item exposure rate frequency distributions were both observed, hence the decrease in mean exposure rate. However, the mean maximum and standard deviation of the exposure rates remained at similar levels. This implied that there were still a substantial number of items with high exposure rates in the two designs. In addition, the mean percentage of items never administered greatly increased for the item-level CAT (to 43% from 14%), showing very poor pool utilization. The MST design maintained good pool utilization as all 253 items used in its panels were consistently administered at least once.

Thus, in summary, shortening the test had a notably positive effect on item exposure and utilization for the testlet-level CAT. It dramatically worsened the pool utilization for the item-level CAT. And it generally decreased the item exposure rates for the MST design, but did not affect its pool utilization.

Table 4.17: Descriptive statistics of testlet exposure rates – short test (short test length, full item pool, normal ability distribution condition)

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.087	0.089	0.301
Item-Level CAT	0.087	0.088	0.301
MST[a]	0.093	0.057	0.285

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 43 (of 46) testlets in its assembled panels

[b] SD: Standard Deviation

Table 4.18: Frequency distribution of testlet exposure rates – short test (short test length, full item pool, normal ability distribution condition)

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	2	2	0
.26-.30	4	4	1
.21-.25	1	1	1
.16-.20	2	2	2
.11-.15	4	4	14
.06-.10	10	10	13
.01-.05	24	24	12
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	46	46	43

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.19: Descriptive statistics of item exposure rates – short test (short test length, full item pool, normal ability distribution condition)

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.021	0.019	0.103
Item-Level CAT	0.021	0.041	0.251
MST[a]	0.083	0.054	0.259

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 253 (of 1,008) items in its assembled panels

[b] SD: Standard Deviation

Table 4.20: Frequency distribution of item exposure rates – short test (short test length, full item pool, normal ability distribution condition)

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.26-.30	0	1	5
.21-.25	0	14	6
.16-.20	0	14	11
.11-.15	1	22	63
.06-.10	115	66	79
.01-.05	891	461	90
Not Admin	1	431	0
Not Admin %	0%	43%	0%
Total Items	1008	1008	253

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Overlap Rates

Tables 4.21 and 4.22 give the testlet and item overlap rates for the *short* test length, full item pool, and normal distribution condition. As expected, the testlet overlap

rates in Table 4.21 were very similar to those for the long test length condition (Table 4.15). The item overlap rates in Table 4.22 decreased for all three conditions. The effect was especially notable for the testlet-level CAT as examinees – similar, different or overall – had mean item overlap rates of less than one item, another desirable property if test security is of concern. However, as with the long test length, differences in item overlap rates found between the three test designs were still likely not large enough to have any practical significance.

Table 4.21: Testlet overlap rates – short test length (short test length, full item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0
Item-Level CAT	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0
MST	0.5	0.0	4.0	0.4	0.0	4.0	0.6	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.22: Item overlap rates – short test length (short test length, full item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.8	0.0	16.0	0.6	0.0	15.2	0.8	0.0	16.0
Item-Level CAT	2.1	0.0	19.7	0.9	0.0	15.7	2.3	0.0	19.7
MST	2.5	0.0	21.0	1.8	0.0	21.0	2.6	0.0	21.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Pool Size

Testlet exposure rates

Tables 4.23 and 4.24 show the descriptive statistics and frequency distribution of the testlet exposure rates for the long test length, *reduced* pool and normal distribution condition. Results in these tables are compared with those in Tables 4.13 and 4.14 to evaluate the impact of reducing the size of the item pool on testlet exposure rates.

In comparing these four tables, it can be seen that the main difference between the full and reduced pool size condition is the increase in mean testlet exposure rates (e.g., to .129 from .087 for the CAT designs). This difference is in line with the change in test-length-to-pool-size ratios between the two conditions (to 4:31 from 4:46). A slight upward shift in the testlet exposure rates was also observed in all three test designs as a result. However, these effects appeared to be similar for the three designs and did not negatively affect pool utilization at the testlet level. Thus, the three designs performed similarly well in controlling testlet exposure under the reduced pool condition.

Table 4.23: Descriptive statistics of testlet exposure rates – reduced pool (long test length, reduced item pool, normal ability distribution condition)

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.129	0.092	0.301
Item-Level CAT	0.129	0.092	0.301
MST[a]	0.133	0.073	0.320

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 30 (of 31) testlets in its assembled panels

[b] SD: Standard Deviation

Table 4.24: Frequency distribution of testlet exposure rates – reduced pool (long test length, reduced item pool, normal ability distribution condition)

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	2	2	1
.26-.30	4	4	1
.21-.25	2	2	3
.16-.20	3	2	5
.11-.15	5	5	9
.06-.10	11	11	6
.01-.05	6	6	5
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	31	31	30

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Item Exposure Rates

Tables 4.25 and 4.26 give the descriptive statistics and frequency distribution of the item exposure rates for the long test length, *reduced* item pool and normal distribution condition. Results in these tables are contrasted with those in Tables 4.13 and 4.14 to evaluate the impact of reducing the pool size on item exposure rates.

As observed at the testlet level, for all three test designs, reducing the pool size increased test-length-to-pool-size ratio, hence proportionally increasing the item exposure rates (Chen, Ankenmann and Spray, 2003). For the two CAT designs, for example, the test-length-to-pool-size ratio increased from 1:24 to 1:18, and the mean item exposure rate therefore increased from .042 to .057. A similar proportional increase was found in the mean item exposure rate for the MST design.

This change in test-length-to-pool-size ratio also manifested itself in an upward shift in the distribution of item exposure rates, as observed in the frequency distributions in Table 4.26. The degree of this shift seemed to be similar for the two CAT designs, but it did not change the maximum item exposure rate. The maximum item exposure rate remained at .21 for the testlet-level CAT and .25 for the item-level CAT. Pool utilization for the testlet-level CAT also remained good as all items, on average, were administered at least once. It improved slightly for the item-level CAT as the number of items never administered went from 14% down to 10% in the reduced item pool condition.

Table 4.25: Descriptive statistics of item exposure rates – reduced pool (long test length, reduced item pool, normal ability distribution condition)

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.057	0.039	0.205
Item-Level CAT	0.057	0.065	0.251
MST[a]	0.112	0.075	0.320

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 375 (of 741) items in its assembled panels

[b] SD: Standard Deviation

Table 4.26: Frequency distribution of item exposure rates – reduced pool (long test length, reduced item pool, normal ability distribution condition)

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	0	0	10
.26-.30	0	5	13
.21-.25	1	39	31
.16-.20	19	29	37
.11-.15	96	61	111
.06-.10	201	153	60
.01-.05	424	378	114
Not Admin	0	76	0
Not Admin %	0%	10%	0%
Total Items	741	741	375

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

For the MST design, the effect of reducing the item pool was more noticeable. The maximum exposure rate, for example, went from .26 with the full item pool to .32

with the reduced pool. Higher proportions of items with high exposure rates were also observed. In the full pool condition, only 31 of the 428 items (about 7%) had exposure rates greater than .20. In the reduced pool, there were 54 out of 375 (about 14%). Pool utilization for the MST design, however, remained good as all item on average were administered at least once. A plausible explanation for the more notable effect on the MST design is that, with the smaller pool, more testlets (and hence items) needed to be used in two modules than with the larger pool. Consequently, reducing the pool size affected the exposure rates for the MST design more than it did for the CAT designs.

Overlap Rates

Tables 4.27 and 4.28 show the testlet and item overlap rates for the long test length, *reduced* item pool, and normal distribution condition. A similar trend observed in the exposure rates was observed for the overlap rates. The trend was that reducing the pool slightly increased the testlet and item overlap rates for all three conditions. The effect was virtually the same for the two CAT designs and was marginally greater for the MST design. The greatest difference observed, in comparing these two tables with Tables 4.15 and 4.16, was for examinees of similar abilities under the MST design. The mean item overlap rate for that group went from 6.1 items with the full item pool (Table 4.16) to 7.2 items with the reduced pool (Table 4.28), a difference of one item on a 42-item test. Thus, although a pool size effect appeared to exist for the overlap rates, it was likely not great enough to be of any practical significance.

Table 4.27: Testlet overlap rates – reduced pool (long test length, reduced item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.8	0.0	4.0	0.7	0.0	4.0	0.9	0.0	4.0
Item-Level CAT	0.8	0.0	4.0	0.6	0.0	4.0	0.9	0.0	4.0
MST	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.28: Item overlap rates – reduced pool (long test length, reduced item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	3.4	0.0	40.2	2.4	0.0	33.6	3.6	0.0	39.3
Item-Level CAT	5.4	0.0	40.7	2.5	0.0	30.7	6.0	0.0	40.7
MST	6.8	0.0	42.0	4.6	0.0	42.0	7.2	0.0	42.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Test Length × Pool Size Interaction

To check for any test length × pool size interaction effects on the exposure properties, the exposure and overlap statistics for the *short* test length, *reduced* pool and normal distribution condition were examined. These results are presented in Tables 4.29 to 4.34. Comparing these results with what was observed for the *long* test length, *reduced* pool and normal distribution condition (in Tables 4.23 to 4.28), a similar test length effect was found in the reduced item pool as it was in the full item pool. Thus, as with measurement effectiveness, there did not appear to be a test length × pool size interaction effect on the exposure control properties for any of the three test designs.

Table 4.29: Descriptive stats of testlet exposure rates (short test length, reduced item pool, normal ability distribution condition)

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.129	0.094	0.301
Item-Level CAT	0.129	0.092	0.301
MST[a]	0.129	0.082	0.296

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used all 31 testlets in its assembled panels

[b] SD: Standard Deviation

Table 4.30: Frequencies of testlet exposure rates (short test length, reduced item pool, normal ability distribution condition)

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	2	2	0
.26-.30	5	4	4
.21-.25	1	1	3
.16-.20	3	3	1
.11-.15	4	4	12
.06-.10	12	12	6
.01-.05	6	5	6
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	31	31	31

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.31: Descriptive stats of item exposure rates (short test length, reduced item pool, normal ability distribution condition)

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.028	0.020	0.113
Item-Level CAT	0.028	0.049	0.250
MST[a]	0.104	0.079	0.294

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 201 (of 741) items in its assembled panels

[b] SD: Standard Deviation

Table 4.32: Frequencies of item exposure rates (short test length, reduced item pool, normal ability distribution condition)

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.31-.35	0	0	1
.26-.30	0	0	16
.21-.25	0	15	18
.16-.20	0	16	3
.11-.15	1	31	52
.06-.10	124	80	43
.01-.05	616	309	69
Not Admin	0	290	0
Not Admin %	0%	39%	0%
Total Items	741	741	201

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.33: Testlet overlap rates (short test length, reduced item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.8	0.0	4.0	0.7	0.0	4.0	0.8	0.0	4.0
Item-Level CAT	0.8	0.0	4.0	0.7	0.0	4.0	0.9	0.0	4.0
MST	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.34: Item overlap rates (short test length, reduced item pool, normal ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.9	0.0	16.7	0.7	0.0	15.6	0.9	0.0	16.7
Item-Level CAT	2.3	0.0	20.1	1.0	0.0	16.1	2.6	0.0	20.1
MST	3.4	0.0	21.0	2.5	0.0	21.0	3.6	0.0	21.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Ability Distribution

Testlet Exposure Rates

Results for the testlet exposure rates for the long test length, full pool and *skewed* distribution condition are presented in Tables 4.35 and 4.36. They are compared with the results in Tables 4.11 and 4.12 to assess the impact of changing the underlying ability distribution from a normal one to a negatively-skewed one on testlet exposure rates.

The results for the two CAT designs were strikingly similar between the two conditions. This implied that at the testlet-level, the CAT designs were relatively robust to this change in the ability distribution. The MST design, however, was affected substantially. Because the test-length-to-pool-size ratio did not change from the normal distribution condition, the MST design maintained the mean testlet exposure rate of .11. However, its mean maximum exposure rate rose sharply from .26 to .41, and its mean standard deviation went from .06 to .11, indicating a much wider distribution of testlet exposure rates. This was reflected in the frequency distribution of the testlet exposure rates where 4 of the 37 (or 11%) testlets on average had testlet exposure rates greater than .31, compared to none in the normal distribution condition. This was also the first instance of the MST design having less than perfect pool utilization. It did not occur in all replications as the average number of testlets never administered was 0.4 (or 1%), but it is also no longer zero like in the other conditions. Thus, changing the underlying ability distribution from a normal one to a negatively-skewed one had a notable differential effect on the MST compared to the two CAT designs.

Table 4.35: Descriptive statistics of testlet exposure rates – skewed distribution (long test length, full item pool, skewed ability distribution condition)

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.087	0.092	0.301
Item-Level CAT	0.087	0.094	0.301
MST[a]	0.108	0.106	0.405

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 37 (of 46) testlets in its assembled panels

[b] SD: Standard Deviation

Table 4.36: Frequency distribution of testlet exposure rates – skewed distribution (long test length, full item pool, skewed ability distribution condition)

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	1
.36-.40	0	0	1
.31-.35	2	2	2
.26-.30	4	5	2
.21-.25	1	0	2
.16-.20	1	1	0
.11-.15	3	3	13
.06-.10	9	8	2
.01-.05	26	27	14
Not Admin	0.0	0.0	0.4
Not Admin %	0%	0%	1%
Total Testlets	46	46	37

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Item Exposure Rates

Tables 4.37 and 4.38 give the descriptive statistics and frequency distribution of the item exposure rates for the long test, full pool and *skewed* distribution condition. These results are compared to those in Tables 4.13 and 4.14 to assess the impact of changing the underlying ability distribution on item exposure rates.

Similar trends observed at the testlet level were also observed at the item level. The CAT designs seemed relatively robust to the change in ability distribution as most of the results were very similar between the normal and skewed conditions. The one notable difference was slightly poorer pool utilization for the item-level CAT, where the percentage of never-administered items rose from 14% to 17%.

For the MST design, the change in underlying distribution had a more obvious effect on the item exposure rates. Just like at the testlet level, a much wider distribution of item exposure rates was observed as the mean maximum item exposure rate grew from .26 to .41 and the mean standard deviation went from .06 to .10. This can also be seen in the item exposure rate frequencies where 31 items (or 7%) on average had testlet exposure rates greater than .31 compared to only one testlet (or 0.2%) in the normal distribution condition. The MST design also had less than perfect pool utilization as, on average, 4 items (or 1%) were never administered to any examinees.

Recall that one of the main reasons this condition was included was to assess the extent to which a MST is impacted by a mismatch between the assumed and actual underlying ability distribution. The MST panels were constructed with the assumption of an underlying normal ability distribution (see, for example, the panel test information functions in Figure 3.7). As these results indicate, the effect was indeed quite different for the MST compared to the two CAT designs.

Table 4.37: Descriptive statistics of item exposure rates – skewed distribution (long test length, full item pool, skewed ability distribution condition)

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.042	0.040	0.205
Item-Level CAT	0.042	0.061	0.251
MST[a]	0.098	0.104	0.405

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 428 (of 1,008) items in its assembled panels

[b] SD: Standard Deviation

Table 4.38: Frequency distribution of item exposure rates – skewed distribution (long test length, full item pool, skewed ability distribution condition)

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	6
.36-.40	0	0	7
.31-.35	0	0	18
.26-.30	0	6	17
.21-.25	2	55	23
.16-.20	23	19	7
.11-.15	92	37	128
.06-.10	147	101	26
.01-.05	744	619	192
Not Admin	0	173	4
Not Admin %	0%	17%	1%
Total Items	1008	1008	428

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Overlap Rates

Tables 4.39 and 4.40 show the testlet and item overlap rates for the long test, full pool, and *skewed* distribution condition. At the testlet level, the overlap rates for all three test designs were similar to what was found in the normal distribution condition. Any differences observed between the two conditions were likely not significant practically.

At the item level, however, the MST design was again more impacted by the change in ability distribution. The item overlap rate went from 5.8 to 8.7 items for all examinees, from 3.8 to 5.1 items for different examinees, and from 6.1 to 9.4 items for similar examinees. These were average increases of 2-3 items on a 42 item test, which may not be great practical concern, except when compared to the CAT designs. The testlet-level CAT, for example, had mean item overlap rates of 3.3 and 3.4 items, nearly identical to the rates in the normal distribution condition. So, on average, the MST design had 5 more overlapping items for all examinees and 6 for examinees of similar ability compared to the testlet-level CAT. This fact, combined with the higher percentage of items with high exposure rates, would pose a security concern for the MST design in a practical setting, especially in light of the robustness of the CAT designs to changes in the underlying ability distribution.

FOR COMPLETE RESULTS

Note that the results presented in this chapter were only a subset of the fully-crossed 24 study conditions (3 test designs × 2 test lengths × 2 pool sizes × 2 ability distributions). Results for any study conditions not presented in this chapter can be found in Appendices B and C.

Table 4.39: Testlet overlap rates – skewed distribution (long test length, full item pool, skewed ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.7	0.0	4.0	0.7	0.0	4.0	0.8	0.0	4.0
Item-Level CAT	0.7	0.0	4.0	0.7	0.0	4.0	0.8	0.0	4.0
MST	0.8	0.0	4.0	0.7	0.0	4.0	1.0	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table 4.40: Item overlap rates – skewed distribution (long test length, full item pool, skewed ability distribution condition)

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	3.3	0.0	35.7	2.7	0.0	32.0	3.4	0.0	35.7
Item-Level CAT	5.4	0.0	41.0	3.0	0.0	31.1	5.9	0.0	41.0
MST	8.7	0.0	42.0	5.1	0.0	42.0	9.4	0.0	42.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

CHAPTER FIVE: DISCUSSION

This chapter provides a discussion of the study results. It includes three main sections. First, the research questions are addressed based on the findings of the study. Conclusions and practical applications of these findings are then described. Limitations of the study and direction for future research are given in the final section.

RESEARCH QUESTIONS

In general, how do the three adaptive test designs compare in their measurement effectiveness and exposure control properties?

Results from the long test length, full item pool and normal ability distribution condition were used to compare the three adaptive test designs at an overall level. This condition was chosen because it was most representative of the test structure for the operational statewide reading assessment from which the study data were drawn.

For measurement effectiveness, the three test designs performed very similarly in terms of measurement accuracy, both overall and across various points on the ability (θ) scale. In terms of measurement precision, the CAT that adapted between and within testlets (item-level CAT) produced the best overall results, followed by the three-stage MST design (MST), and then the CAT that adapted between testlets only (testlet-level CAT). The differences in overall measurement precision were small and likely not of any practical significance.

The measurement precision at various points on the θ scale, however, was notably different for the three test designs. For the two CAT designs, measurement precision across the θ scale was a direct reflection of the shape of the test information function for the item pool (see Figure 3.1). Thus, the CAT designs measured examinees with lower ability (θ values around -1) more precisely than examinees with moderate to high θ

values. In contrast, for the MST design, the measurement precision across the θ scale mirrored the target test information functions (TIFs) to which the modules and panels were constructed. As such, the most precisely measured examinees were those with θ values around zero while examinees with θ values at the two ends of the scale were measured with less precision. The consequence of these test design characteristics was that the item-level CAT had the best measurement precision across the θ scale, while the MST had comparable precision measuring examinees with higher ability. At the low end of the θ scale, however, the measurement precision of the MST was worse than both CAT designs.

In terms of exposure control properties, the three designs performed well and similarly at the *testlet* level, with generally low exposure rates and good testlet utilization. At the *item* level, the testlet-level CAT produced the best result in terms of exposure control and pool utilization. The item-level CAT had relatively worse pool utilization, with an average of 14% of the items never administered to any examinees. It was, however, able to maintain the maximum exposure control rate specified by the progressive-restrictive procedure and the exposure rates were generally low for items that were administered. The MST design had excellent pool utilization, but the considerably higher percentage of items with high exposure rates could be a test security concern. This was likely an artifact of the fixed number of paths an examinee could take though each panel and the fact that only a subset of items from the full item pool was included in the MST panels. So, in reality, it had a smaller pool of items available for administration. The three test designs performed well in terms of test overlap and any differences between the designs were likely not of any practical significance.

These findings support results from several previous studies comparing CAT and MST (Kim & Plake, 1993; Luecht, Nungester & Hadadi, 1996, Schnipke & Reese, 1997; Patsula, 1999; Jodoin, 2003; Hambleton & Xing, 2006). All these studies found that *item-based* CATs yielded better measurement precision than *item-based* MSTs. This study confirmed the general findings, but did so in the context of *testlet-based* CATs and MSTs – the testlet-based item-level CAT produced better measurement precision than the testlet-based MST design. The current research also found that, across the ability scale, the measurement precision of the two types of designs was characterized by different aspects of the test structure. The CAT was dependent on the distribution of test information its item pool, while the MST reflected the TIFs to which it was constructed.

Looking specifically at studies that incorporated exposure control procedures into their CAT and MST comparisons, Patsula (1999) and Jodoin (2003) found that CATs with *conditional* exposure control procedures performed better than MSTs in both measurement precision and exposure control properties. Davis and Dodd (2003), on the other hand, found that CATs performed slightly worse than MSTs when they implemented a *randomization* exposure control procedure. The CAT designs in this study implemented the progressive-restrictive exposure control procedure, a hybrid procedure with both randomization and conditional components (Revuelta & Pondosa, 1998). As such, perhaps not surprisingly, the results appeared to strike a middle ground between these previously findings. The testlet-level CAT produced the best exposure control properties, but it also had the worst measurement precision when compared to the other two designs. The item-level CAT had the best measurement precision, but this was traded off with less than ideal pool utilization. The MST had more consistent measurement precision across the θ scale, but it also had a higher percentage of items with high exposure rates. Thus, it is difficult to declare any test design as better based on

these findings. Each design appears to have its strengths and weakness, and may be the ideal design a given testing context.

Finally, the results for the testlet-level CAT design were strikingly similar to those found for the analogous condition in Boyd's (2003) study. This was encouraging because the testlet-level CAT was a direct extension of Boyd's (2003) progressive-restrictive (maximum exposure rate = .30) condition under the TRT model. Thus, these results cross-validate her findings on a completely different dataset.

Does the total test length have a differential effect on the measurement effectiveness and exposure control properties of the three adaptive test designs?

In terms of measurement effectiveness, shortening the test from 42 items to 21 items substantially decreased the measurement precision of all three adaptive test designs. This decrease in measurement accuracy was most prevalent at the ends of the θ scales while the decrease in precision was consistently found across the scale.

The item-level CAT, however, seemed more robust to this test length effect, as its measurement precision did not decrease to the same degrees as the other two designs. This was likely due to the additional level of flexibility afforded by adaptively selecting items within testlets. While the testlet-level CAT and MST designs were tied to the five or six items pre-selected with each testlet, the item-level CAT was still able to choose the five or six items that were ideal for each examinee conditional on the estimated θ value. As a result, the measurement precision of the item-level CAT was less affected by the shorter test length.

In terms of exposure control properties, the test length had a clear differential effect on the three test designs at the item level. It resulted in a near ideal situation for the testlet-level CAT with consistently low item exposure rates and good pool utilization. It dramatically worsened the pool utilization for the item-level CAT. And it marginally

decreased the item exposure rates for the MST design, but did not affect its pool utilization.

For the testlet-level CAT, the positive effect was probably due to the fact that, under the short test length condition, far more five- or six- item permutations (with very similar psychometric properties) were available for selection with each testlet in the pool. Thus, the exposure rates for these items were spread out more evenly among the testlet items, keeping the exposure rates low and relatively homogeneous.

For the item-level CAT, this result was a classic trade-off between measurement precision and pool utilization. Being able to adaptively choose items within a testlet implied that the probability an item would be administered was strongly related to its psychometric properties. Those that provide less item information across the θ scale were administered sparingly. And with a shorter test, more items of this type had virtually no chance at all to be administered, hence the dramatic increase in proportion of items never administered observed for the item-level CAT.

For the MST design, shorter tests meant that fewer items from the original item pool were included in its modules and panels. However, once items were included, their exposure rates were more a function of the test structure (such as number of panels and the between-stage routing properties) than they were the actual test length. Thus, the effect of test length on the MST's exposure control properties was not as obvious.

The general finding of shorter tests having less measurement precision is a well-known property in classical test theory. Thus, the overall results come as no surprise. None of the studies involving CAT and MST, however, have investigated the differential effects of test length on the properties of each test design. Thus, the results of this study provide a starting point for this area of research. Also, these study findings provide partial support for the conjecture made by Stark and Chernyshenko (2006) that test length

does indeed have a substantial impact on the estimation ability of a MST. However, whether this effect is greater than any of the other MST design considerations, as suggested by Stark and Chernyshenko (2006), remains a subject for further research.

Are the adaptive test designs affected differently by a reduction in the test-length-to-pool-size ratio in terms of their measurement and exposure control properties?

In terms of measurement effectiveness, the study results showed that none of the three adaptive test designs were notable affected by the reduction in their available item pool. Nor were any test length × pool size interaction effect observed.

In terms of exposure control properties, reducing the pool size did have a more discernable effect on the exposure rates for the MST design than it did the two CAT designs. A plausible explanation for this is that, with the smaller pool, more testlets and items needed to be used in multiple modules than were with the larger pool. Using a testlet or item in multiple modules effectively multiplied the chance it could be administered. Thus, having more testlets and items in multiple modules would notably shift the frequency distribution of their exposure rates upwards. No test length × pool size interaction, however, were observed in the MST design, nor were they found in the two CAT designs.

The findings in this condition were somewhat unexpected. It was originally hypothesized that measurement precision and exposure control rates would be adversely affected by a reduction of available items. However, this effect was not apparent in the study results. One possible explanation for this is that reducing the item pool size to two-thirds of the original pool size was not dramatic enough to affect the properties of the test designs. The CAT design still had plenty of available items and testlets in the pool to choose from, while the MST design still had sufficient items on which to build its modules and panels. Given that no test length × pool size interaction effect was found, it

would be interesting to see whether reducing the available item pool to half the original size would have lead to a more discernable effect.

What effect does a mismatch between actual and assumed underlying ability distribution have on the measurement and exposure control properties of the MST? How does this effect compare to those of the two CAT designs?

In this study, a mismatch between the actual and assumed underlying ability distribution for the MST was simulated by generating θ values from a negatively-skewed distribution while building the MST modules based on a normality assumption for the θ values. Comparing the results between the normal and skewed ability distribution conditions helped answer this research question.

In the terms of measurement effectiveness, the study found that negatively-skewing the ability distribution decreased the overall measurement accuracy and precision of all three test designs. This was due in large part to the higher proportion of θ values in the high range of the ability scales that were poorly estimated. The decrease in measurement precision, however, was smaller for the MST design. This was likely due to the fact that, in the negatively-skewed distribution, there was a substantially smaller portion of examinees with lower θ values. This happened to be the range on the θ scale for which the MST had the worse measurement precision, relative to the two CAT designs. Thus, the absence of these examinees off-set the degree of measurement precision loss in the MST when compared to the two CAT designs.

In terms of exposure control properties, the exposure control properties of the CAT designs were relatively robust to the change in underlying distribution. This was expected because the CAT algorithms made no inherent assumptions about the underlying ability distribution. Thus, as long as the item pool was sufficient large (as

was the case in the study), the change in ability distribution should have little effect on the frequency distribution of the exposure rates.

In stark contrast, negatively-skewing the ability distribution had a substantial effect on the MST. The distribution of testlet and item exposure rates became much wider and the maximum exposure rates rose sharply, indicating a considerable imbalance in the proportion of times testlet and items were exposed to examinees. This was likely a reflection of the mismatch of the test candidates' ability levels and the item pool. The testlet and item overlap statistics also increased substantially, indicating that examinees have a lot more of their tests in common with one another. Such a scenario would pose a security concern for the MST design in a practical setting, especially if examinees can share test items with one another between test administrations.

CONCLUSIONS AND PRACTICAL APPLICATIONS

As the movement towards computer-based testing continues to move forward in large-scale educational assessments, understanding the properties of various computer-based test designs, such as CAT and MST, becomes evermore important. High-stakes decisions are and will continue to be made based on these assessment results. As such, it is vitally important that the psychometric properties of the test designs are well-known to ensure the defensibility of the testing program. The test components investigated in this dissertation, such as test length, item pool size, assumed examinee proficiency, the use of testlets and TRT, and exposure control procedures, are all issues and decision points that practitioners in the field face regularly. Thus, the findings from this study contribute to the expanding knowledge base in this field of research and provide practical guidelines to programs that are considering CAT or MST as a test design.

First, this study demonstrates the viability of using the 3PL-TRT as a measurement model for testlet-based adaptive tests. Previous to this, Boyd (2003) was

the only study that examined the use of TRT in testlet-based CATs. The measurement effectiveness results in this study are very similar to those in Boyd's study. They are also comparable to the results found in studies using one of the polytomous IRT models to measure testlet-based CATs (e.g. Pastor, Dodd & Chang, 2003; Davis, Pastor, Dodd, Chiang & Fitzpatrick, 2003; Davis & Dodd, 2003; Boyd, 2003; Davis, 2004). This is also the first study that uses TRT as the measurement model for a testlet-based MST. And the generally comparable results of the MST to that of the CAT design demonstrate that the TRT model is a viable option for testing programs considering the MST design.

Next, this current research shows the effectiveness of the progressive-restrictive procedure (Revuelta & Pondosa, 1998) at controlling item and testlet exposure rates while ensuring good measurement precision in a testlet-based CAT. The findings cross-validate the results from Boyd's (2003) progressive-restrictive procedure condition for the traditional testlet-level CAT. The findings also show that the procedure effectively controls item exposure rates for the testlet-based item-level CAT. Thus, testing programs that administer testlet-based CATs can consider implementing the progressive-restrictive procedure to help bolster test security without jeopardizing measurement precision.

Last, this study informs researchers and practitioners about the properties as well as advantages and disadvantages of the testlet-based CAT and MST designs. It is the first study to explore the use of the testlet-based item-level CAT, a fulfillment of what Wainer, Bradlow and Du (2000) termed *ad hoc testlet construction*. This method of CAT administration is only possible with the use of the 3PL-TRT and the study finds that it is able to achieve improved measurement precision over the traditional testlet-level CAT. Thus, if a testing program has an item pool in which each testlet has substantially more items available than are actually administered, then the use of a testlet-based item-level CAT can improve measurement precision over the traditional testlet-level CAT. It

should be noted, however, that the measurement precision achieved by the testlet-level CAT is already considered very good, especially in light of Boyd's (2003) findings. Thus, the improved precision afforded by the item-level CAT may not be practically significant in cases where the test is reasonably long, such as the 42-item test condition in this study. However, if a testing program desires a substantially shorter test (for example, around 20 items), due to either concerns over examinee fatigue or a limited item pool, then the study findings show that the improvement in precision is likely sufficiently significant to warrant the implementation of a testlet-based item-level CAT design.

Even if a testing program chooses to use the traditional testlet-level CAT, the current research demonstrates that by pre-specifying different permutations of each testlet in the pool, good pool utilization and consistently low item exposure rates can be achieved with only a moderate level of measurement precision loss.

The MST design generally does not achieve the same level of measurement precision as the testlet-based item-level CAT, nor does it have exposure control properties that are as good as the traditional testlet-based CAT. However, its results are comparable to both CAT designs and the advantage of greater administrative control in the test development process makes it an attractive and viable alternative to CAT. Caution though should be taken in constructing the MST modules and panels so that it is being built to the correct underlying ability distribution. This can be done through a thorough investigation of the testing population prior to the initial test administration and regular monitoring of the score distributions in subsequent administrations.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

While the findings in this dissertation provide answers to the posed research question, it also raises additional questions due to limitations in the study design. These limitations can serve as starting points for future research in this area.

First, the 3PL-TRT was the only measurement model used in this study. Boyd (2003) found that the partial credit model (Masters, 1982) generally yielded more accurate and precise ability estimates than the 3PL-TRT in testlet-based CAT simulations. Her study, however, only included the traditional testlet-level CAT design. The use of a polytomous IRT model, such as the partial credit model, is not possible with the testlet-based item-level CAT implemented in this study. However, given that the item-level CAT yielded the best measurement precision in this study, it would be interesting to compare its performance to the traditional testlet-level CAT measured using the partial credit model. This would represent a comparison of the best case scenarios for the TRT and polytomous IRT models in the context of testlet-based CATs. Future studies can also compare the use of the 3PL-TRT model with the polytomous IRT model in measuring testlet-based MSTs.

Also, while the use of a testlet-based item-level CAT does lead to a gain in measurement precision, these benefits need to be weighed against the non-psychometric considerations in building a testlet. For example, content experts may argue that adaptively selecting items to include with a testlet solely based on psychometric properties does not appropriately account for context effects or satisfy content specifications for the testlet. To address this concern, content balancing procedures, such as Kingsbury and Zara (1989), could be included as part of the within-testlet item selection algorithm. However, including content balancing procedures may negatively affect the measurement precision gained by using an item-level CAT. Future studies can therefore examine the effects of item-level content balancing procedures on the measurement properties of a testlet-based item-level CAT.

As stated earlier, the fact that reducing the item pool had virtually no effect on the measurement and exposure control properties of any of the test designs seems counter-

intuitive. A plausible explanation for this finding is that the reduced item pool used in this study was not small enough to make any meaningful impact. Thus, future studies can explore the use of even smaller item pools to see whether item pool size truly makes a difference and, if so, whether its effect differs for the various adaptive test designs.

The MST design used in this study represents only one of many ways to implement a MST. Thus, caution should be taken to not over-interpret or over-generalize the results from this one specific MST design. Future studies can include additional MST designs that differ in, for example, stage structure, routing methods, and test assembly method, and see how they compare with the one implemented in this study and with the CAT designs.

Finally, the implementation of either testlet-based CAT or MST designs may, in practice, be limited primarily to low-stakes testing situations. This is not because of any severe shortcomings in these test designs. As shown in this and many previous studies, CAT and MST designs have very desirable psychometric properties. The greatest challenge moving forward for adaptive test designs, however, is likely in the area of score reporting. It is usually difficult for exam stakeholders such as teachers, students, and parents, to interpret the results from an adaptive test. For example, two students taking an adaptive test can get the same percentage of items correct on each of their tests, but end up with very different ability estimates and hence, very different reported scores and mastery classifications. In a high-stakes testing environment, score misinterpretation can often lead to erroneous reports in the media and potential legal challenges to the testing program. Thus, future studies should examine ways of reporting adaptive tests results that are more easily understood by the general public, thereby enhancing the defensibility of the testing program.

APPENDICES

APPENDIX A: DISTRIBUTION OF PARAMETERS IN FULL ITEM POOL

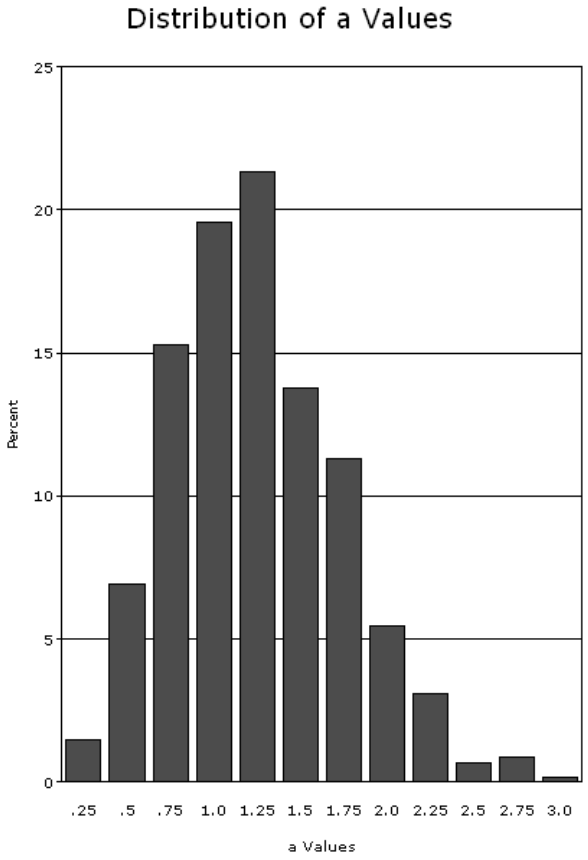

Figure A.1: Distribution of discrimination (a) parameters in the item pool

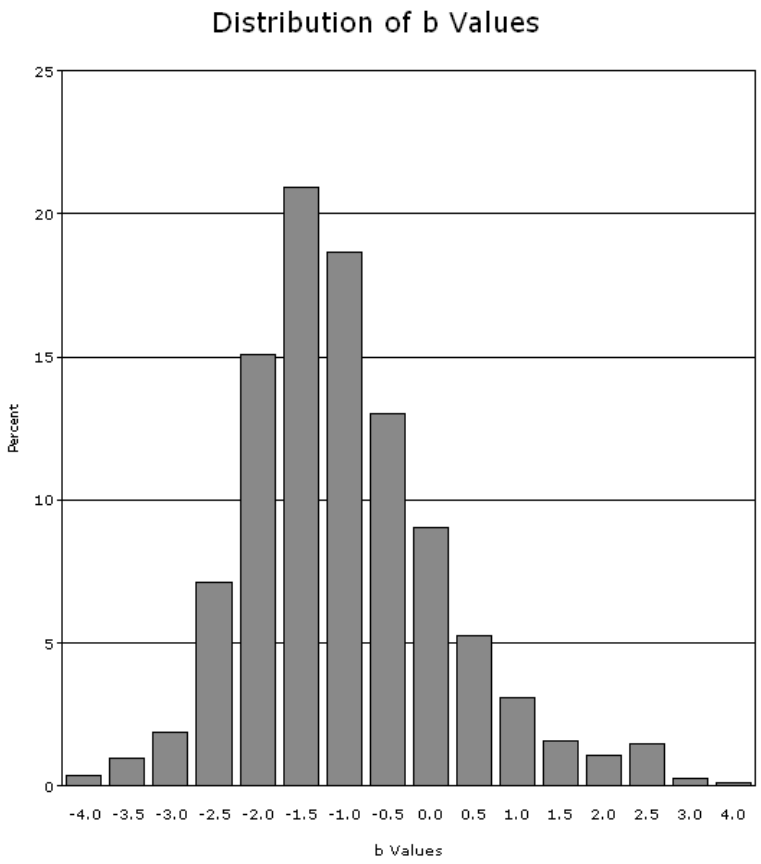

Figure A.2: Distribution of difficulty (b) parameters in the item pool

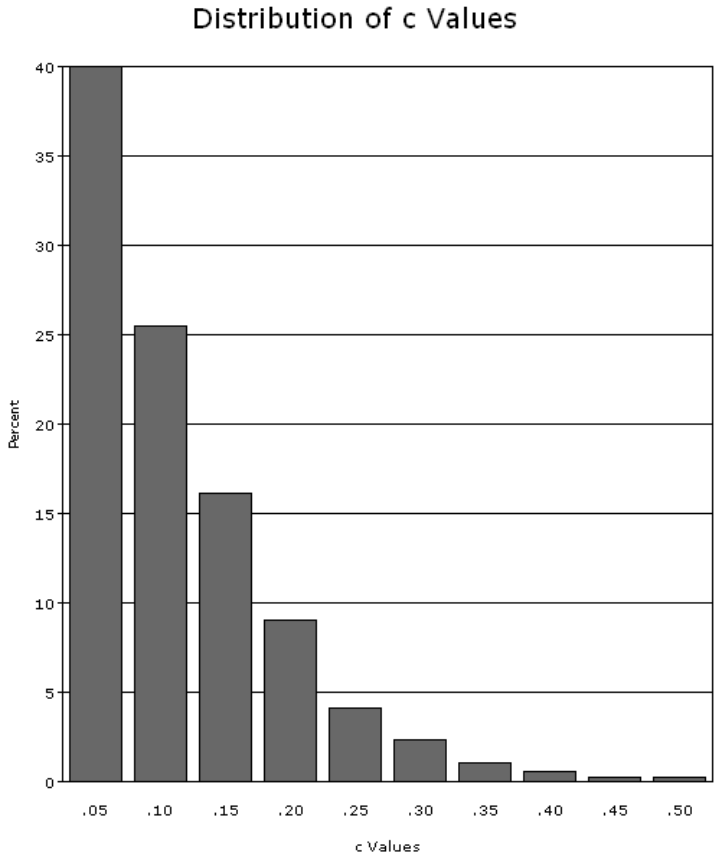

Figure A.3: Distribution of pseudo-guessing (c) parameters in the pool

APPENDIX B: MEASUREMENT EFFECTIVENESS – OTHER CONDITIONS

Long Test Length, Reduced Item Pool, Skewed Distribution

Table B.1: Descriptive stats of the estimated θ - long, reduced, skewed

	Estimated θ		
Test Design	Grand Mean (Min, Max)	Mean SE (Min, Max)	Mean Correlation (Min, Max)
Testlet-Level CAT	1.133 (1.082, 1.169)	0.482 (0.474, 0.489)	0.887 (0.878, 0.895)
Item-Level CAT	1.226 (1.183, 1.256)	0.386 (0.382, 0.39)	0.916 (0.909, 0.922)
MST	1.218 (1.172, 1.267)	0.403 (0.398, 0.409)	0.914 (0.905, 0.921)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

Table B.2: Bias, RMSE and AAD of the estimated θ - long, reduced, skewed

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.365 (0.328, 0.385)	0.595 (0.563, 0.612)	0.479 (0.459, 0.489)
Item-Level CAT	0.271 (0.241, 0.292)	0.488 (0.47, 0.506)	0.391 (0.375, 0.405)
MST	0.279 (0.247, 0.294)	0.496 (0.475, 0.52)	0.400 (0.382, 0.42)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

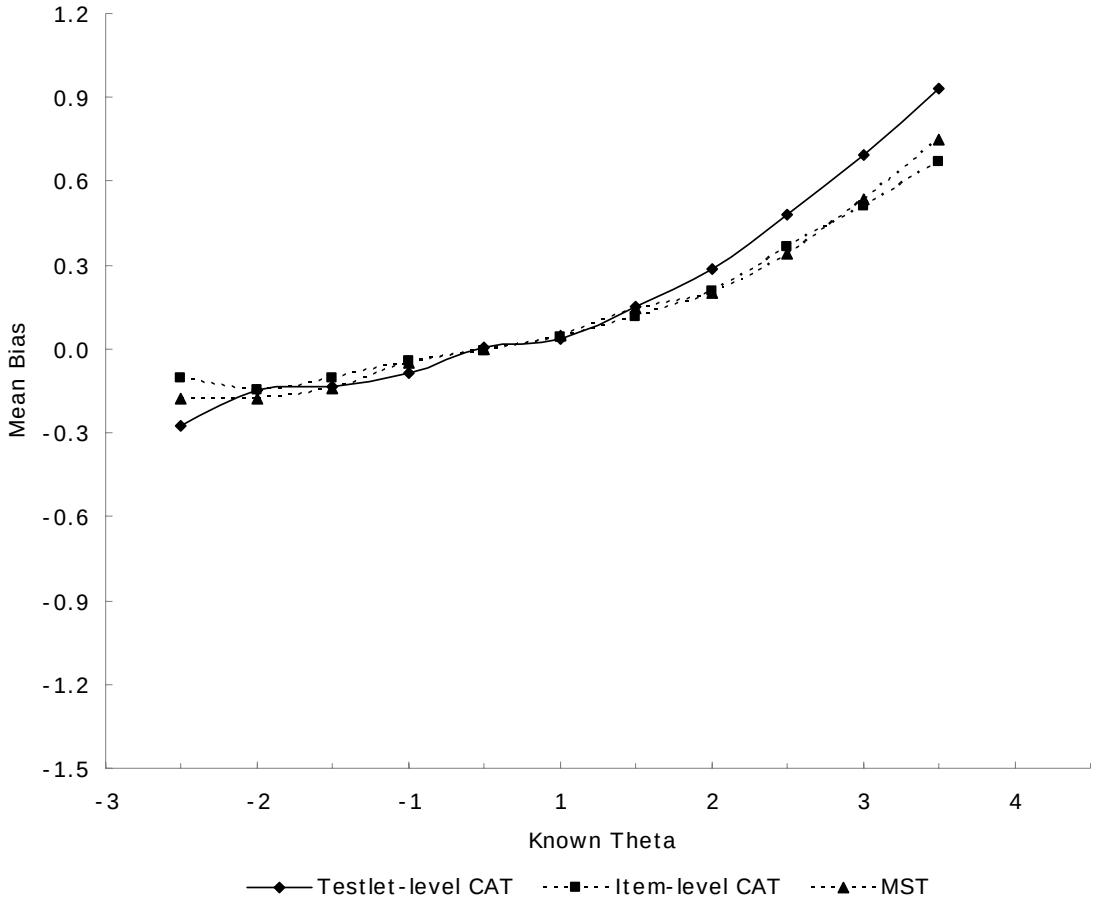

Figure B.1: Conditional mean bias plot – long, reduced, skewed

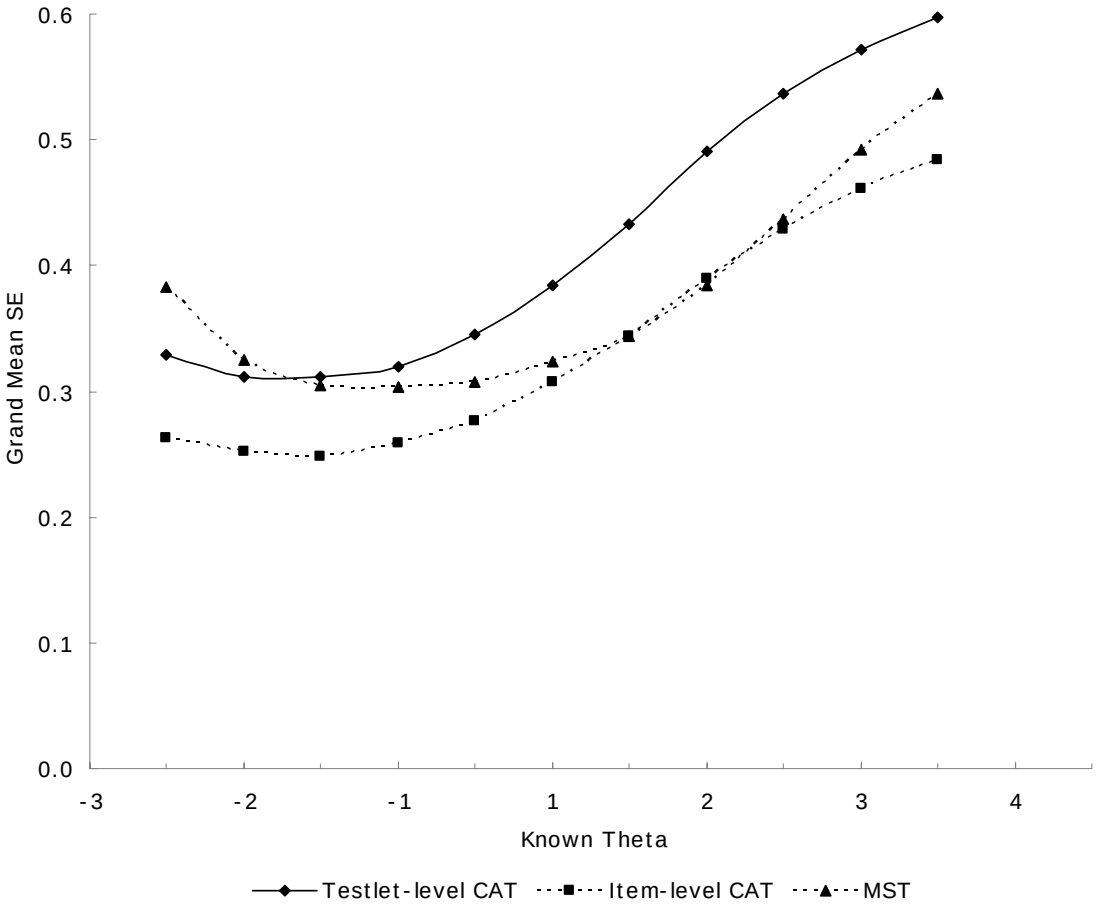

Figure B.2: Conditional grand mean standard error plot – long, reduced, skewed

Short Test Length, Full Item Pool, Skewed Distribution

Table B.3: Descriptive stats of the estimated θ - short, full, skewed

Test Design	Estimated θ[a] Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	0.961 (0.908, 1.005)	0.648 (0.639, 0.652)	0.817 (0.799, 0.831)
Item-Level CAT	1.226 (1.183, 1.256)	0.386 (0.382, 0.39)	0.916 (0.909, 0.922)
MST	1.087 (1.042, 1.126)	0.533 (0.527, 0.537)	0.866 (0.854, 0.875)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = 1.497, min mean = 1.426, max mean = 1.548

[b] SE: standard error; Correlation: between known and estimated θ's

Table B.4: Bias, RMSE and AAD of the estimated θ - short, full, skewed

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.536 (0.5, 0.557)	0.797 (0.784, 0.814)	0.656 (0.645, 0.676)
Item-Level CAT	0.271 (0.241, 0.292)	0.488 (0.47, 0.506)	0.391 (0.375, 0.405)
MST	0.411 (0.373, 0.433)	0.652 (0.621, 0.672)	0.533 (0.503, 0.552)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

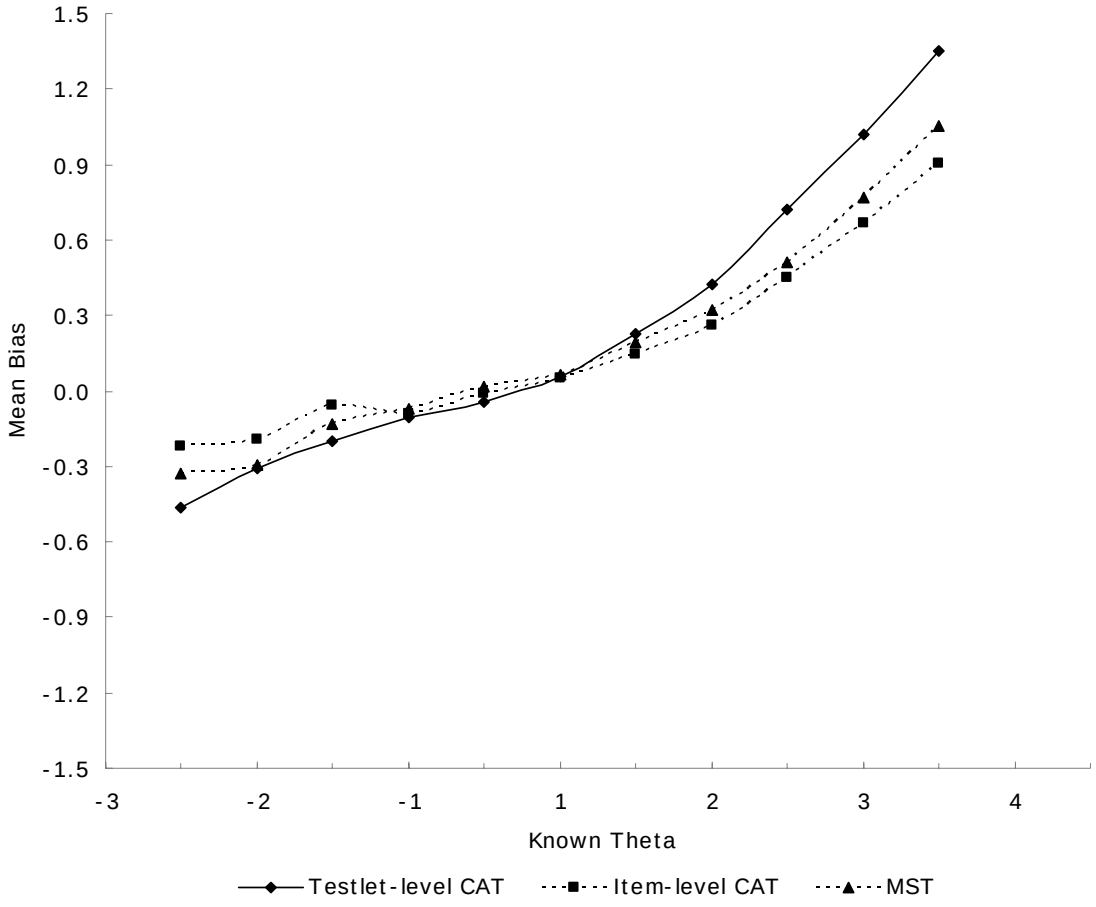

Figure B.3: Conditional mean bias plot – short, full, skewed

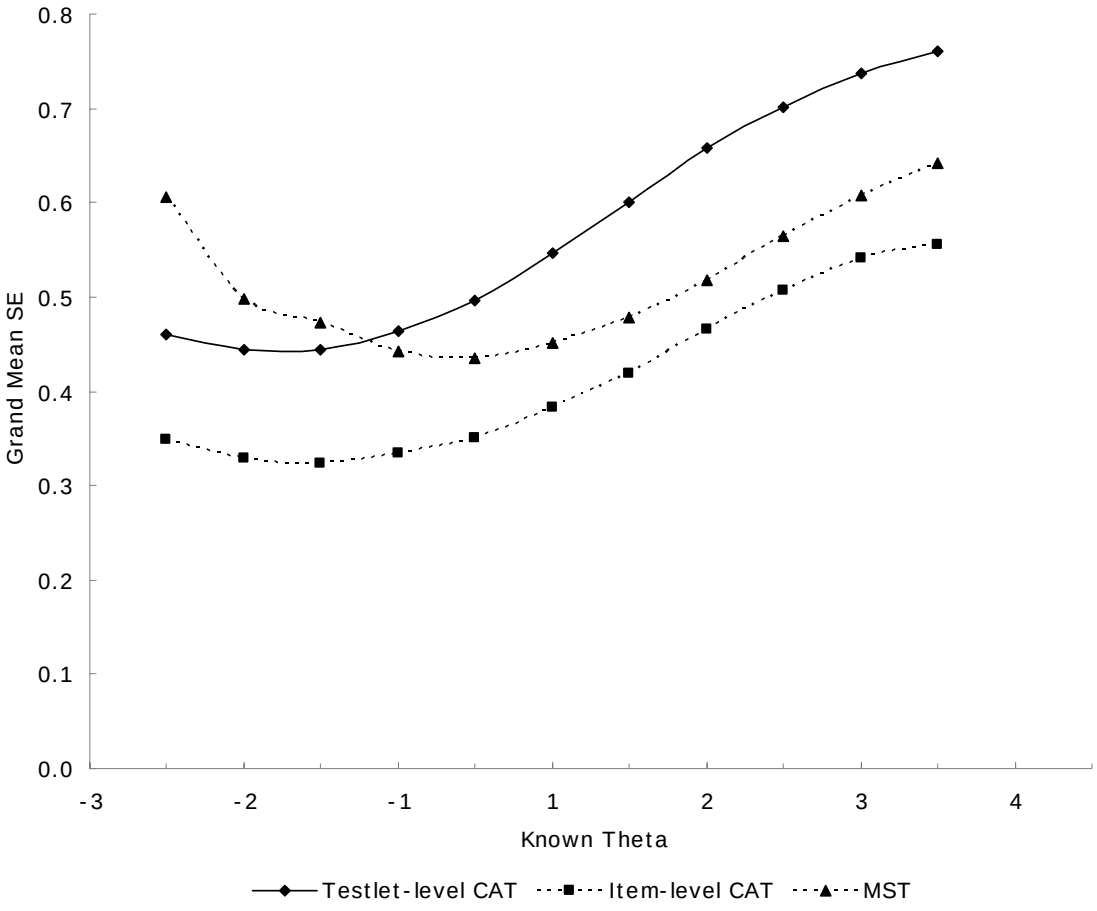

Figure B.4: Conditional grand mean standard error plot – short, full, skewed

Short Test Length, Reduced Item Pool, Skewed Distribution

Table B.5: Descriptive stats of the estimated θ - short, reduced, skewed

Test Design	Estimated θ[a] Grand Mean (Min, Max)	Mean SE[b] (Min, Max)	Mean Correlation[b] (Min, Max)
Testlet-Level CAT	0.961 (0.919, 0.997)	0.645 (0.637, 0.652)	0.821 (0.807, 0.841)
Item-Level CAT	1.154 (1.106, 1.203)	0.463 (0.458, 0.468)	0.885 (0.876, 0.897)
MST	1.087 (1.039, 1.122)	0.533 (0.527, 0.536)	0.865 (0.847, 0.871)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

[a] Known θ's: grand mean = 1.497, min mean = 1.426, max mean = 1.548

[b] SE: standard error; Correlation: between known and estimated θ's

Table B.6: Bias, RMSE and AAD of the estimated θ - short, reduced, skewed

Test Design	Bias (Min, Max)	RMSE (Min, Max)	AAD (Min, Max)
Testlet-Level CAT	0.537 (0.504, 0.571)	0.792 (0.754, 0.82)	0.654 (0.618, 0.685)
Item-Level CAT	0.343 (0.302, 0.382)	0.581 (0.563, 0.598)	0.468 (0.454, 0.487)
MST	0.410 (0.363, 0.44)	0.653 (0.616, 0.676)	0.535 (0.501, 0.561)

Note: All statistics were computed from across 10 replications; each replication contained 1,000 observations

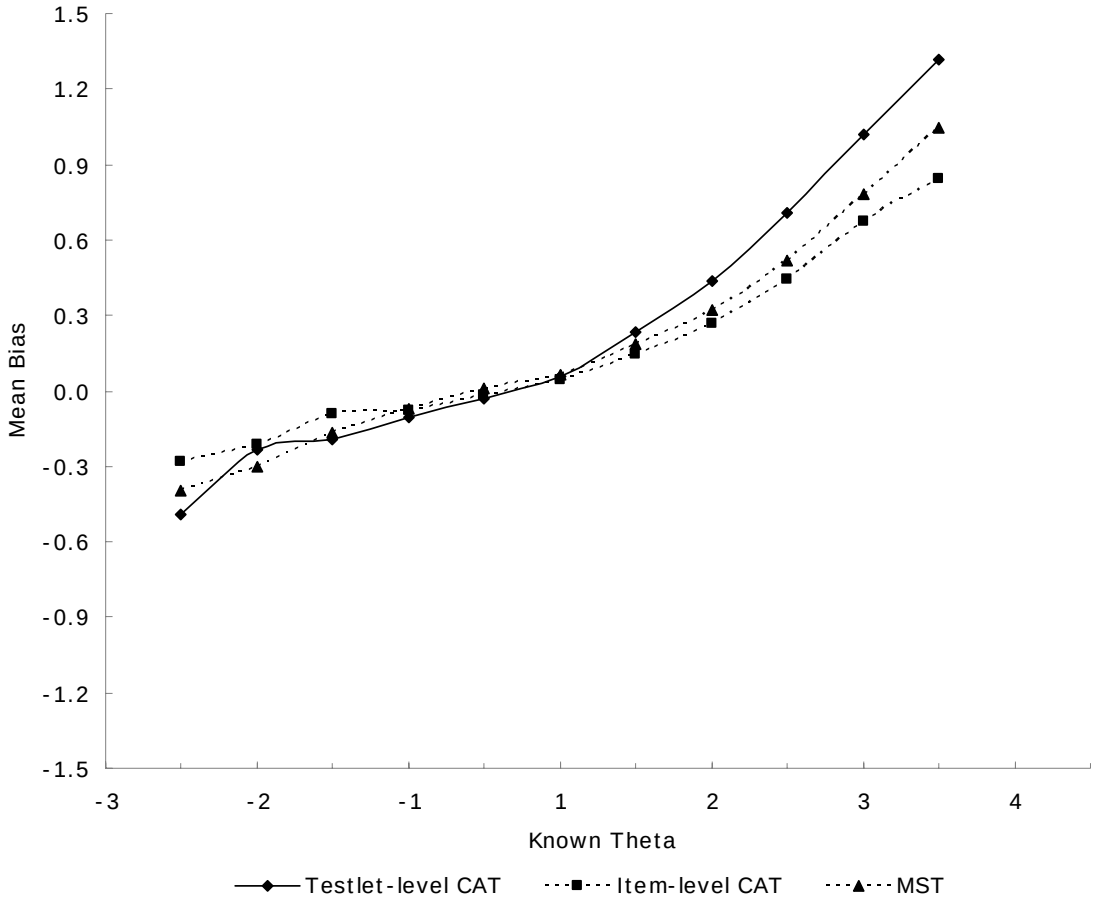

Figure B.5: Conditional mean bias plot – short, reduced, skewed

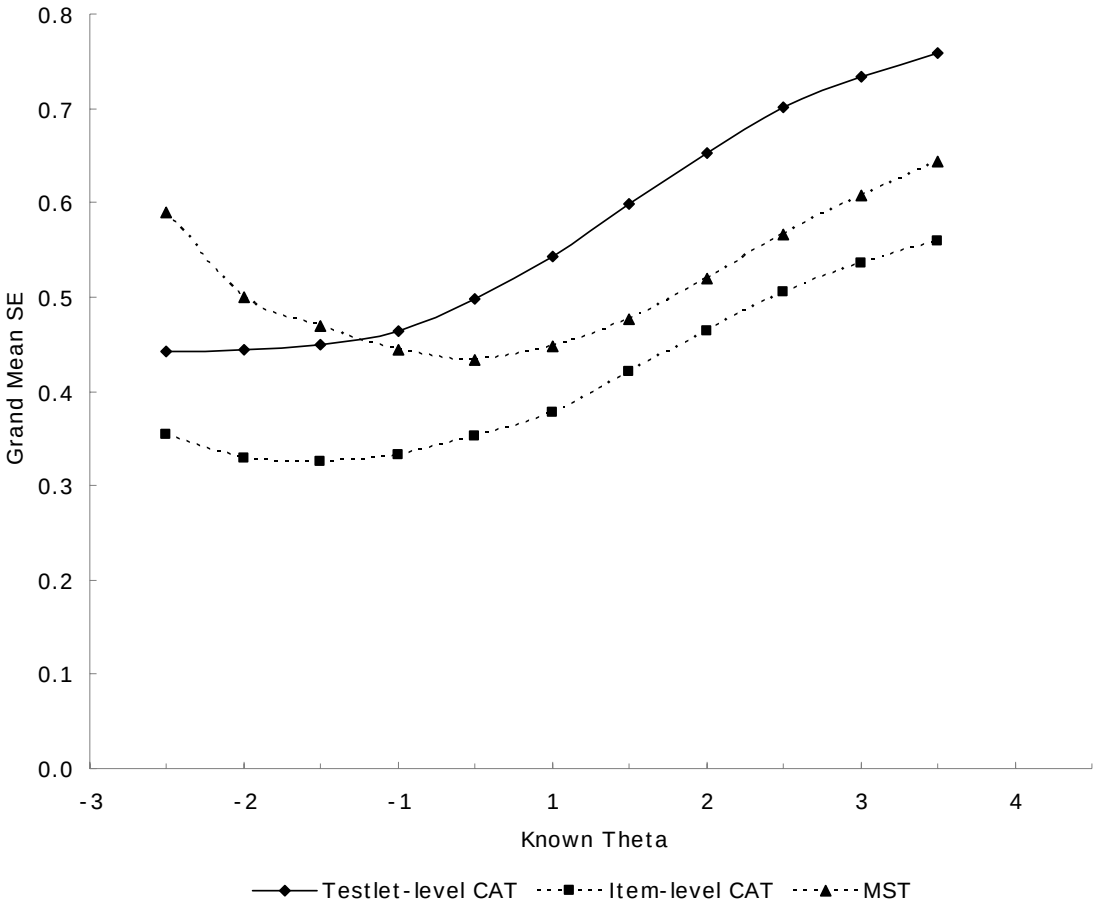

Figure B.6: Conditional grand mean standard error plot – short, reduced, skewed

APPENDIX C: EXPOSURE CONTROL – OTHER CONDITIONS

Long Test, Reduced Item Pool, Skewed Distribution

Table C.1: Descriptive stats of testlet exposure rates – long, reduced, skewed

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.129	0.097	0.301
Item-Level CAT	0.129	0.098	0.301
MST[a]	0.133	0.116	0.405

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 30 (of 31) testlets in its assembled panels

[b] SD: Standard Deviation

Table C.2: Descriptive stats of item exposure rates – long, reduced, skewed

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.057	0.041	0.210
Item-Level CAT	0.057	0.069	0.251
MST[a]	0.112	0.114	0.405

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 375 (of 741) items in its assembled panels

[b] SD: Standard Deviation

Table C.3: Frequencies of testlet exposure rates – long, reduced, skewed

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	1
.36-.40	0	0	1
.31-.35	2	2	2
.26-.30	4	5	3
.21-.25	1	0	3
.16-.20	1	1	0
.11-.15	6	5	9
.06-.10	10	11	2
.01-.05	7	6	10
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	31	31	30

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.4: Frequencies of item exposure rates – long, reduced, skewed

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	7
.36-.40	0	0	6
.31-.35	0	0	18
.26-.30	0	9	32
.21-.25	2	52	27
.16-.20	23	24	8
.11-.15	101	50	89
.06-.10	180	149	24
.01-.05	436	350	162
Not Admin	0	108	3
Not Admin %	0%	15%	1%
Total Items	741	741	375

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.5: Testlet overlap rates – long, reduced, skewed

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.8	0.0	4.0	0.8	0.0	4.0	0.8	0.0	4.0
Item-Level CAT	0.8	0.0	4.0	0.7	0.0	4.0	0.9	0.0	4.0
MST	0.9	0.0	4.0	0.8	0.0	4.0	1.0	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.6: Item overlap rates – long, reduced, skewed

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	3.6	0.0	37.5	3.1	0.0	33.2	3.7	0.0	37.4
Item-Level CAT	5.9	0.0	41.0	3.5	0.0	31.5	6.3	0.0	41.0
MST	9.5	0.0	42.0	6.0	0.0	42.0	10.2	0.0	42.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Short Test, Full Item Pool, Skewed Distribution

Table C.7: Descriptive stats of testlet exposure rates – short, full, skewed

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.087	0.091	0.301
Item-Level CAT	0.129	0.098	0.301
MST[a]	0.093	0.089	0.311

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 43 (of 46) testlets in its assembled panels

[b] SD: Standard Deviation

Table C.8: Descriptive stats of item exposure rates – short, full, skewed

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.021	0.020	0.096
Item-Level CAT	0.057	0.069	0.251
MST[a]	0.083	0.086	0.311

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 253 (of 1,008) items in its assembled panels

[b] SD: Standard Deviation

Table C.9: Frequencies of testlet exposure rates – short, full, skewed

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	0
.36-.40	0	0	0
.31-.35	2	1	1
.26-.30	5	5	3
.21-.25	1	1	1
.16-.20	1	1	4
.11-.15	4	3	11
.06-.10	9	10	2
.01-.05	25	26	21
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	46	46	43

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.10: Frequencies of item exposure rates – short, full, skewed

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	0
.36-.40	0	0	0
.31-.35	0	0	5
.26-.30	0	2	15
.21-.25	0	18	6
.16-.20	0	13	21
.11-.15	0	23	55
.06-.10	124	52	14
.01-.05	883	415	138
Not Admin	1	485	1
Not Admin %	0%	48%	0%
Total Items	1008	1008	254

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.11: Testlet overlap rates – short, full, skewed

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.7	0.0	4.0	0.7	0.0	4.0	0.7	0.0	4.0
Item-Level CAT	0.7	0.0	4.0	0.7	0.0	4.0	0.8	0.0	4.0
MST	0.7	0.0	4.0	0.6	0.0	4.0	0.8	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.12: Item overlap rates – short, full, skewed

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.8	0.0	16.0	0.7	0.0	14.4	0.8	0.0	16.0
Item-Level CAT	2.4	0.0	20.5	1.3	0.0	15.2	2.6	0.0	20.5
MST	3.6	0.0	21.0	2.2	0.0	21.0	3.9	0.0	21.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Short Test, Reduced Item Pool, Skewed Distribution

Table C.13: Descriptive stats of testlet exposure rates – short, reduced, skewed

	Testlet Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.129	0.095	0.301
Item-Level CAT	0.129	0.097	0.301
MST[a]	0.129	0.105	0.317

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used all 31 testlets in its assembled panels

[b] SD: Standard Deviation

Table C.14: Descriptive stats of item exposure rates – short, reduced, skewed

	Item Exposure Rate		
Test Design	Grand Mean	Mean SD[b]	Mean Maximum
Testlet-Level CAT	0.028	0.020	0.098
Item-Level CAT	0.028	0.052	0.251
MST[a]	0.104	0.103	0.317

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

[a] MST design used only 201 (of 741) items in its assembled panels

[b] SD: Standard Deviation

Table C.15: Frequencies of testlet exposure rates – short, reduced, skewed

Testlet Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	0
.36-.40	0	0	0
.31-.35	2	2	1
.26-.30	4	5	5
.21-.25	1	1	3
.16-.20	1	1	3
.11-.15	6	7	5
.06-.10	10	10	3
.01-.05	6	7	11
Not Admin	0	0	0
Not Admin %	0%	0%	0%
Total Testlets	31	31	31

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.16: Frequencies of item exposure rates – short, reduced, skewed

Item Exposure Rate	Testlet-level CAT	Item-level CAT	MST
.41-.50	0	0	0
.36-.40	0	0	0
.31-.35	0	0	7
.26-.30	0	1	24
.21-.25	0	19	14
.16-.20	0	16	16
.11-.15	0	31	26
.06-.10	128	73	20
.01-.05	613	271	93
Not Admin	0	331	1
Not Admin %	0%	45%	0%
Total Items	741	741	201

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.17: Testlet overlap rates – short, reduced, skewed

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.8	0.0	4.0	0.8	0.0	4.0	0.8	0.0	4.0
Item-Level CAT	0.8	0.0	4.0	0.8	0.0	4.0	0.8	0.0	4.0
MST	0.8	0.0	4.0	0.7	0.0	4.0	0.9	0.0	4.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

Table C.18: Item overlap rates – short, reduced, skewed

Test Design	Overall			Different Ability			Similar Ability		
	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean	Grand Mean	Min Mean	Max Mean
Testlet-Level CAT	0.9	0.0	17.0	0.8	0.0	16.0	0.9	0.0	17.0
Item-Level CAT	2.6	0.0	21.0	1.5	0.0	15.9	2.8	0.0	21.0
MST	4.3	0.0	21.0	2.9	0.0	21.0	4.6	0.0	21.0

Note: All statistics were averaged across 10 replications; each replication contained 1,000 observations

REFERENCES

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence*. ACT research report series, 87-14. Iowa City, IA: American College Testing.

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement*, *27*, 241–253.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669-679.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal Testlet Pool Assembly for Multistage Testing Design. *Applied Psychological Measurement*, *30(3)*, 204-215.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, *28*, 147–164.

Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes' essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*, 293–315.

Bergstrom B. A., & Lunz, M. E. (1992). Confidence in pass/fail decisions for computer adaptive and paper-and-pencil examinations. *Evaluation and the Health Professions*, *15(4)*, 453-464.

Bergstrom B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates.

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing*, *5(3)*, 319-330.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems*. Unpublished doctoral dissertation, University of Texas at Austin.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries.* (Research Report No. 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Chang, H. H. (2004). Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences*, (pp. 117-133). Sage Publications.

Chang, H. H., Qian, J., & Ying, Z. (2001). A-stratified multistage CAT with b blocking. *Applied Psychological Measurement*, *25*, 333-341.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229.

Chang, S. W., Ansley, T. N., & Lin, S. H. (2000). *Performance of item exposure control methods in computerized adaptive testing: Further explorations.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Chen, S., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40(2)*, 129-145.

Chen, S., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation on CAT using the partial credit model. *Educational and Psychological Measurement*, *53*, 61-77.

Chen, S., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computer adaptive testing using the rating scale model. *Educational and Psychological Measurement*, *57(3)*, 422-439.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich College.

Cronbach, L. J., & Glaser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Chuah, S. C., Drasgow, F., & Luecht, R. M. (2006). How big is big enough? Sample size requirements for CAST item parameter estimations. *Applied Measurement in Education*, *19(3)*, 241-255.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. Mills, M. T. Potenza, J. J. Fremer and W. C. Ward (Eds.), *Computer-Based Testing: Building the Foundation for Future Assessments*, Mahwah, NJ: Lawrence Erlbaum Associates.

Davey T., & Parshall, C. G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, *28(3)*, 165-185.

Davis, L. L., & Dodd, B. G. (2003). Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT. *Applied Psychological Measurement*, *27(5)*, 335-356.

Davis, L. L., Pastor, D. A., Dodd, B. G., Chiang, C., & Fitzpatrick, S. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, *4(1)*, 24-42.

Dodd, B. G., De Ayala, R. J., & Koch W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19 (1)*, 5-22.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.), *Computerized adaptive testing: A primer (2nd ed.)*. Mahwah, NH: Lawrence Erlbaum Associates.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: estimation procedures, population distribution, and the item pool characteristics. *Applied Psychological Measurement*, *29(6)*, 433-456.

Gulliksen, H. (1950). *Theory of Mental Tests*. Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H, & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage Publications.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, *19(3)*, 221-239.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26(2)*, 44-52.

Hasting, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *54*, 93-108.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B. K. Waters, and J. R. McBride (Eds.), *Computerized adaptive testing-from inquiry to operation* (pp. 141-144). Washington, D.C.: American Psychological Association.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: application to psychological measurement*. Homewood, IL: Dow-Jones Irwin.

Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pools*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19(3)*, 203-220.

Kim, H., & Plake, B. S. (1993). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

Kingsbury, G. G. (1990). Adapting adaptive testing: using the MicroCAT testing in a local school district. *Educational Measurement: Issues and Practice*, *9(2)*, 3-6.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 257-283). New York: Academic Press.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359-375.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, *4(3)*, 241-261.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: an exploratory IRT model fit tool. *Applied Psychological Measurement*, *9*, 281-288.

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computer adaptive testing using partial credit scoring. *Applied Measurement in Education*, *2(4)*, 335-357.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: methods and practices (2nd ed.)*. New York: Springer-Verlag.

Leung, C., Chang, H. H., & Hau, K. (1999). *An enhanced stratified computerized adaptive testing design*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*, 367–386.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, *36*, 227–242.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test*. Educational Testing Service RB-74–30.

Lord, F. M. (1977). A broad-range test of verbal ability. *Applied Psychological Measurement*, *1*, 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loyd, B. H. (1984). *Efficiency and precision in two-stage adaptive testing*. Paper presented at the Annual Meeting of the Eastern Educational Research Association, West Palm Beach, Florida.

Luecht, R. M. (1998). CASTISEL [Computer Software]. Philadelphia, PA: National Board of Medical Examiners.

Luecht, R. M. (2000). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2003). *Exposure control using adaptive multistage item bundles*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19(3)*, 189-202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229-249.

Luecht, R. M., Nungester, R. J., & Hadidi, A. (1996). *Heuristic-based CAT: Balancing item information, content and exposure*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, *31*, 251-263.

Lunz, M. E., & Stahl, J. A. (1998). *Patterns of item exposure using a randomized CAT algorithm.* Paper presented at the Annual meeting of the National Council on Measurement in Education, San Diego, CA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp.223-226). New York, Academic Press.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19(3)*, 185-187.

Metropolis, N., Rosenblith, A. W., Rosenblith, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examination general test. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 117-135). Mahwah, NJ: Lawrence Erlbaum Associates.

Mills, C. N., & Stocking, M. L. (1995). *Practical issues in large-scale high-stakes computerized adaptive testing* (Research Report 95-23). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: item and test scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59-71.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Parshall, C. G., Davey, T., & Nering, M. L. (1998). *Test development exposure control for adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Parshall, C. G., Hogarty, K. Y., & Kromrey, J. D. (1999). *Item exposure in adaptive tests: an empirical investigation of control strategies*. Paper presented at the annual meeting of the Psychometrics Society, Lawrence, KS.

Parshall, C. G., Spray, J., Kalohn, J., Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Pastor, D.A., Dodd, B.G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, *26 (2)*, 147-163.

Patsula, L. (1999). *Comparison of computerized adaptive testing and multi-stage testing*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Reckase, M. D. (1981). Final *report: procedures for criterion referenced tailored testing*. Columbia: University of Missouri, Educational Psychology Department.

Reckase, M. D. (1989). Adaptive testing: the evolution of a good idea. *Educational Measurement: Issues and Practice*, *18(3)*, 11-15.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21(1)*, 25-36.

Revuelta, J., & Ponsoda, V. (1996). Metodos sencillos para el control de las tasas de exposicion en tests adaptativos informatizados [Simple methods for item exposure control in CATs]. *Psicologica*, *17*, 161- 172.

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35(4)*, 311-327.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53(3)*, 349-360.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement*, *12*, 397-409.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Schnipke, D. L., & Reese, L. M. (1997). *A comparison of testlet-based test designs for computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Segall D. O., & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Service Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Bunchanan (Eds.), *Innovations in computerized assessment* (pp. 35-65). Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28(3)*, p.237-247.

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: widely or narrowly applicable? *Applied Measurement in Education*, *19(3)*, 257-260.

Stocking, M. L., & Lewis, C. (1995). *A new method for controlling item exposure in computer adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing, *Journal of Educational and Behavioral Statistics*, *23(1)*, 57-75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 163-182). Netherlands: Kluwer Academic Publishers.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17(3)*, 277-292.

Stocking, M. L., Swanson, L., & Pearlman, M. (1991). *Automatic item selection (AIS) methods in the ETS testing environment* (Research Memorandum 91-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). The application of an automated item selection method to real data. *Applied Psychological Measurement*, *17(2)*, 167-176.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17(2)*, 151-166.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of the Military Testing Association. San Diego: Navy Personnel Research and Development Center.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528-550.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: a primer* (pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, *26*, p.247-260.

Thomasson, G. L. (1998). *CAT item exposure control: new evaluation tools, alternate methods and integration into a total CAT program.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego.

Tuerlinckx, F., and De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181-195.

Urry, V. W. (1977). Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement*, *14*, 181-196.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195–211.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized adaptive testing: theory and practice* (pp. 1-25). Netherlands: Kluwer Academic Publishers.

Vispoel, W. P. (1998). Reviewing and changing answers on computer adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*, 328–345.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*, 191-208.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Psychological Measurement*, *8 (2)*, 157-187.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized adaptive testing: theory and practice* (pp. 245-269). Netherlands: Kluwer Academic Publishers.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: a primer* (pp. 271-299). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, *29*, 243–251.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-202.

Wainer H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: a comparison of hierarchical and linear structures. Journal of Educational Measurement, *28(4)*, 311-323.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and profiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: a primer* (pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339-368.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence of reliability? *Educational Measurement: Issues and Practice*, *16*, 22-29.

Wang, X., Bradlow, E. T., & Wainer, H. (2000). A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement*, *26(1)*, 109-128.

Wang, X., Bradlow, E. T., & Wainer, H. (2001). The SCORIGHT computer program [Computer program]. Princeton, NJ: Educational Testing Service.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17(4)*, 17-27.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report No. 73-3). Minneapolis, MN: University of Minnesota, Department of Psychology. Psychometric Methods Program.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wright, B.D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Yi, Q., & Chang, H. H. (2000). *Multiple stratification CAT designs with content control*. Unpublished manuscript.

Zara, A. R. (1989). A research proposal for field testing CAT for nursing licensure examinations. In *Delegate Assembly Book of Reports 1989*. Chicago: National Council of State Boards of Nursing, Inc.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

VITA

Leslie Keng was born in Toronto, Ontario, Canada on January 23, 1974, the son of Gene C. Keng and Liza W. Keng. He completed his late-elementary school and middle school education in Taiwan, R.O.C. and his high school education at John F. Ross CVI in Guelph, Ontario, Canada. In 1993, he began his undergraduate studies at the University of Waterloo, Waterloo, Ontario. To fulfill program requirements and obtain his teaching certificate in Ontario, he also attended teacher's college at Queen's University, Kingston, Ontario. He received the degrees of Bachelor of Mathematics in computer science from the University of Waterloo and Bachelor of Education from Queen's University in June 1998. From 1998 to 2002, Leslie worked as a technical trainer for Trilogy in Austin, Texas. In August 2002, he entered the Graduate School at the University of Texas at Austin. He received his Master of Science in statistics in August 2004. He worked as a psychometric intern for Pearson in Austin, Texas from June to December in 2005 and then from June to December in 2006. He began working full time for Pearson as a research associate in July, 2007.

Permanent address: 8001 Tantara Court, Austin, Texas, 78729

This dissertation was typed by the author.